Does Context Matter in Lexical Simplification?

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

Abstract

Lexical Simplification is the task of replacing a word in a sentence in order to make the sentence easier to read. Starting with sentences from Wikipedia, we collected a data set of possible substitutions using the Amazon Mechanical Turk service. In this paper we investigate whether the surrounding sentence, or "the context", is important in making a good substitution.

1 Introduction

The task of Lexical Simplification is a piece of the much larger project of full text simplification. Text simplification can help children, second language learners, and people with cognitive disabilities, among others, read complex texts. It is also an important tool in preprocessing texts for NLP tasks. Much work has been done in the field but there still remains fundamental questions.

In this paper we investigate the role of the surrounding context of a target word in the task of Lexical Simplification. We started with a sentence aligned corpus of English Wikipedia and Simple English Wikipedia. After selecting a small number of candidate words based on certain criterion we solicited possible substitutions in a context provided environment and a context withheld environment from the Amazon Mechanical Turk service.

Using this data we compared the environments using multiple metrics. Based on multiple indicators, indeed context does have an effect on the substitutions. We also use several psycholinguistic metrics to further explain the discrepancies amongst the substitutions.

2 Related Work

some stuff here

3 The Data Set

The candidate words and the corresponding contexts were taken from a sentence aligned corpus of English Wikipedia(Normal) and Simple English Wikipedia(Simple). We then used the Giza word alignment to generate Normal-Simple candidate word pairs. A Normal-Simple candidate pair consists of a sentence, the index of the candidate word in that sentence, the aligned word from the Simple corpus, and the parts of speech of both those words. These pairs were then filtered to 27,000 pairs using the following criteria:

- The aligned words must have the same part of speech.
- The part of speech must be noun, verb, adjective, or adverb.
- The normal word must not be on the Dale-Chall or Ogden lists of simple words.
- The simple worst must have a higher corpus frequency than the normal word.

From this set of 27,000 we chose 24 with the only discrimination being that the sentences be unique among those 24.

4 First Experiment

4.1 Set-Up

To run our experiments we utilized the Amazon Mechanical Turk system(mTurk). mTurk is an online marketplace for human intelligence tasks, or HITs. We posted a Normal-Simple pair as a HIT in two different environments:

- The full sentence is displayed. The candidate is red and bold. The annotator is asked to provide a simple substitute for the candidate word in a text box. We will call this the context environment.

- The candidate word is displayed without the sentence. The part of speech and a definition from WordNet are displayed next to the candidate word. The annotator is asked to provide a simple substitute for the candidate word in a text box. We will call this the no-context environment.

We gathered 50 annotations for each Normal-Simple pair in each environment. In total we gathered 240 annotations. We rewarded an annotator $0.03 for each annotation they submitted. In total we had NUM annotators complete annotations. The most annotations received from one annotator was NUM and the least was NUM. The mean received from one annotator was NUM. An annotator was not allowed to submit more than one annotation per candidate word pair per environment.

4.2 Analysis

We investigated the annotations using several different metrics. There are three general types comparisons we made.

- Between the two sets of context annotations. We refer to this as context comparison.

- Between the two sets of no-context annotations. We refer to this as no-context comparison.

- Between one set of context annotations and one set of no-context annotations. We refer to this as basic comparison.

candidate word	no-context	context	difference
satellite	0.4132	0.1548	0.2584
settled	0.6843	0.4360	0.2483
longevity	0.7535	0.5243	0.2292
intensified	0.7572	0.5382	0.2190
practice	0.7993	0.5940	0.2054
predominantly	0.4429	0.2791	0.1639
footballer	0.4715	0.3159	0.1556
founded	0.4315	0.2782	0.1533
express	0.4912	0.3560	0.1352
ancient	0.2145	0.0925	0.1220
retired	0.6308	0.5126	0.1183
edited	0.5350	0.4246	0.1104
obligated	0.4150	0.3338	0.0812
carcass	0.4991	0.4244	0.0747
invaded	0.5038	0.4708	0.0330
conceals	0.0580	0.0250	0.0330
produced	0.3334	0.3231	0.0103
sovereign	0.4483	0.4453	0.0030
combat	0.1171	0.1172	0.00
published	0.3784	0.3992	-0.0208
antagonist	0.4069	0.4499	-0.0430
received	0.3227	0.3958	-0.0730
competitions	0.3378	0.4713	-0.1335
installment	0.3889	0.5358	-0.1469

Table 1: Measures of entropy listed in order of difference

Frequency

Entropy

We treated all of the annotations from a single environment for a candidate word as a random variable, X, and calculated the Shannon entropy, $H(X)$ where

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i)$$

We chose $b = 50$ so that $H(X) \in [0, 1]$. The Pearson product-moment correlation coefficient for the data in Table 1 is NUM and the Spearman rank correlation coefficient is NUM.

To comment on the candidate words' entropy we needed to establish a baseline to compare them against. We randomly split every candidate word's annotations for each environment

Distribution	-2σ	Mean	$+2\sigma$
entropy, context	0.0891	0.4131	0.7370
entropy, no-context	0.1042	0.4976	0.8910
$\overline{\text{entropy}}$, context	0.2296	0.3797	0.5299
$\overline{\text{entropy}}$, no-context	0.4640	0.5648	0.6656
entropy Δ, context	-0.5764	0.0834	0.7431
entropy Δ, no-context	-0.5602	0.0842	0.7286
entropy Δ, basic	NUM	NUM	NUM
$\overline{\text{entropy}}$ Δ, context	0.0014	0.1218	0.2421
$\overline{\text{entropy}}$ Δ, no-context	-0.0607	0.0774	0.2155
$\overline{\text{entropy}}$ Δ, basic	NUM	NUM	NUM
Pearson, context	0.7411	0.8346	0.9281
Pearson, no-context	0.7926	0.8682	0.9439
Pearson, basic	0.5939	0.7074	0.8208
Spearman, context	0.6254	0.7627	0.9000
Spearman, no-context	0.6834	0.7999	0.9163
Spearman, basic	0.4109	0.5754	0.7400

Table 2: Distributions using entropy, $\overline{\text{entropy}}$ indicates mean entropy

	entropy, context	$\overline{\text{entropy}}$, context
entropy, no-context	48.89	NULL
$\overline{\text{entropy}}$, no-context	NULL	36.94
	entropy Δ, context	entropy Δ, no-context
entropy Δ, basic	NUM	NUM
entropy Δ, no-context	0.8507	NULL
	$\overline{\text{entropy}}$ Δ, context	$\overline{\text{entropy}}$, no-context
$\overline{\text{entropy}}$ Δ, basic	NUM	NUM
$\overline{\text{entropy}}$ Δ, no-context	18.07	NULL
	Pearson, context	Pearson, no-context
Pearson, basic	125.0789	165.4629
Pearson, no-context	36.5729	
	Spearman, context	Spearman, no-context
Spearman, basic	152.4510	189.3431
Spearman, no-context	32.9698	

Table 3: Absolute value of Welch's t-test

into two sets of 25 annotations. The Shannon entropy was then calculated for each set of 25 annotations.

The first four distributions are straightforward records of entropy for every candidate word and the mean entropy over all candidate words using one random set of 25 annotations to determine a candidate word's entropy.

Similar distributions were generated using the difference between a candidate word's difference in entropy using context, no-context, and basic comparisons. The same was done for the mean entropy over all the candidate words.

We then generated a distribution that measured the Pearson coefficient for context, no-context, and basic comparisons. The same was done for the Spearman coefficient. The results are summarized in Tables 2 and 3.

Similarity

Similarity between a candidate word's context annotations and no-context annotations was measured using cosine similarity and an overlap coefficient. If we let C be a candidate words context annotations and N the no-context annotations then the overlap coefficient is given by

$$\frac{|C \cap N|}{min(|C|, |N|)}$$

If we consider C_i to be the frequency of the ith annotation among all the annotation $\in C$ such that N_i is the frequency of that same annotation among all annotations $\in C$ then the cosine similarity is given by

$$\frac{\sum_{i=1}^{n} C_i \times N_i}{\sqrt{\sum_{i=1}^{n} C_i^2} \times \sqrt{\sum_{i=1}^{n} N_1^2}}$$

The Pearson coefficient for the data in Table 4 is NUM and the Spearman coefficient is NUM. We generated baseline numbers for our similarity metrics much like we did for entropy.

5 Second Experiment

much like First Experiment but less details

6 Outlier Candidate Words

7 Results and Discussion

interpret t values from Table 3.

8 Conclusion

References

candidate word	cosine	overlap coeff.
retired	0.7416	0.1042
installment	0.4613	0.3542
longevity	0.7096	0.3600
intensified	0.6863	0.3958
practice	0.8754	0.4000
sovereign	0.5146	0.4490
settled	0.8386	0.4583
express	0.7334	0.4600
antagonist	0.6662	0.4694
published	0.8390	0.4898
obligated	0.7095	0.5000
footballer	0.7289	0.5600
edited	0.8264	0.5714
carcass	0.8557	0.5714
received	0.9934	0.6200
predominantly	0.9336	0.6600
founded	0.8994	0.6735
invaded	0.9578	0.6800
competitions	0.9653	0.6800
produced	0.9595	0.7000
satellite	0.9933	0.7143
ancient	0.9998	0.8600
conceals	0.9990	0.9400
combat	1.000	0.9400

Table 4: Measures of similarity in order of overlap