An Examination of a Selection of English Lexical Simplification Systems

Joshua Gluck
Computer Science Honors Major / 500 College Avenue
Department of Computer Science / Swarthmore, PA
Swarthmore College / 19081
jgluck2@swarthmore.edu

Abstract

This paper presents the methods and systems I developed to perform the first task (English Lexical Simplification) of SemEval 2012. This task involved ranking synonyms based on their simplicity in a provided context sentence. English Lexical Simplification (ELS) has a number of useful applications such as providing texts for children, the mentally handicapped, and new English speakers. My system relies on n-grams frequencies computed from the Simple English Wikipedia version, ranking each substitution by decreasing frequency of use, as well as n-gram frequencies computed from the British National Corpus (BNC), also ranked by decreasing frequency of use. I also experimented with systems based on word senses, ranking by increasing and decreasing word sense count and word sense entropy according to WordNet, under the theory that the number and distribution of meanings of a substitution has some bearing on whether an individual can easily understand that word in context. My final system, and by far the simplest, involved ranking words by increasing word length, under the theory that shorter words are easier to understand. I achieved a 0.4686 score with the first system, and a 0.4657 score with the second. Due to the smaller size of the Simple English Wikipedia corpus, this suggests that using supervised corpora to calculate frequency may be a key tool for ELS. Ranking by decreasing word sense count, decreasing word sense entropy, and increasing word length all gave similar scores in the low 0.2-0.3 range. This suggests that while word senses and word length may be useful tools for ranking lexical simplicity, they cannot be used effectively individually

1 Introduction

Lexical Simplification consists of determining which synonym in a set of synonyms is the simplest in a given context. It is similar in many respects to the task of Lexical Substitution (McCarthy and Navigli, 2007) in that it involves elements of ranking preference on the basis of a central predetermined criterion (simplicity in the current case), as well as sensitivity to context.

Lexical Simplication envisions a human target audience, and can greatly benefit children, new language learners, people with cognitive disabilities, and in general process information for the sake of making it easier to disseminate to wider audiences which would not be able to understand more complex versions of the material. I used a number of methods based on features that I considered might be linked to lexical simplicity. These include:

Using N-Gram frequencies generated from the Simple English Wikipedia and the British National Corpus (BNC)

Using the number and entropy of word senses, taken from the Wordnet database

Using the length of the words as a ranking criteria

1.1 Task Setup

The English Lexical Simplification shared task at SemEval 2012 (Specia et al., 2012)required systems to rank a number of candidate substitutes (which were provided beforehand) based on their simplicity of usage in a given context. Additionally, the part of speech of the original word in the context was given. For example:

<corpus lang="english">
<lexelt item="bright.a">
<instance id="1">
<context>During the siege , George Robertson had appointed Shuja-ul-Mulk , who was a <head¿bright</head> boy only 12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral .</context> </instance>

In this case, the word is an adjective, a form of the word 'bright', and the potential substitutes to rank are 'intelligent; bright; clever; smart;' Ties between substitute rankings were permitted, and every substitute had to be ranked.

1.2 Corpora

Two corpora were provided: the trial corpus, with 300 examples for developing and fine-tuning systems, and the test corpus, consisting of 1710 for evaluation. These examples, in both corpora, included substitutions for nouns, verbs, adjectives, and adverbs. (although there were relatively few adverbs) Initially, the gold standard for the trial corpus was also given. For the prior example, the gold standard was 'Sentence 1 rankings: intelligent clever smart bright'. This gold standard was generated by new English Speakers' hand annotating the data, and the gold standard for the test corpus, while not initially provided, was evaluated the same way.

The Semeval organizers also provided three baseline systems for lexical simplification: the first a simple randomization of the substitute list, the second one keeps the substitute list in the form provided to participants, and the third relies on frequency rankings based on the Google Web IT corpus.

1.3 Evaluation

The Semeval organizers also provided an evaluation scheme based on pair-wise comparisons between the gold standards output and a given systems output for each example. This system was normalized to produce results from -1 to 1, with -1 being the worst possible ranking scheme for every example, and 1 being the best possible ranking scheme for every example, according to the hand annotated gold standard. I utilized this evaluation scheme to measure the effectiveness of my systems in comparison to the baselines and systems that were provided by participants in SemEval 2012. In addition, to this scheme, I also chose to evaluate my systems based solely on whether their 'simplest' substitute in each example matches the 'simplest' substitute in the gold standard. I did this under the theory that the overall goal of ELS is to provide the simplest sentences possible, and that some systems may be more or less effective at choosing the 'simplest' substitute compared to ordering the less simple substitutes correctly.

1.4 Related Works

Lexical Simplification hasn't received an enormous amount of interest in the NLP community, particularly when compared to Syntactic Simplification. However, with the addition of the papers accepted by the organizers of SemEval 2012 to prior papers on Lexical Simplification, there are certainly enough papers to discuss trends and differences in the field. "For the Sake of Simplicity" (Yatskar et al., 2010) is one example of papers studying lexical simplification prior to the SemEval task. This paper illustrates an attempt to extract word/phrase level lexical simplifications from the Simple English Wikipedia by examining edit histories of pages as well as their corresponding Meta Data. In contrast, the work done by Balder and Moens in 'Text Simplification for Children' (De Belder and Moens, 2010) examines an entire text, and by treating the problem as an integer linear programming problem attempts to make changes that make the entire text as simple to understand as possible rather than making individual phrases or words simpler to understand. The work done by Or Biran and others in "Putting it Simply: a Context-Aware Approach to Lexical Simplification"

(Biran et al., 2011) is closest to the task at hand, insofar as it examines potential systems for choosing the simplest words from a list of potential words in the context of a sentence or short paragraph.

In addition to having differing focuses in terms of the extent of what should be considered for determining simplicity (individual words, the entire document, sentences/paragraph), the NLP community does not necessarily agree on what 'simple' means, with some authors choosing children as their target user groups (De Belder and Moens, 2010), versus people with low literacy levels (Aluisio et al., 2008) and aphasic readers. (Carroll et al., 1998 1999) The target user group for this paper is new English speakers, as that is the group that created the 'gold standards of simplicity' for the first SemEval 2012 task. However, none of the measures implemented in this paper are based on ideas that would solely apply to this group, and so the results may be extendable to others.

2 Preprocessing

2.1 Corpus Constitution

None of the Lexical Simplification methods that I implemented used machine-learning techniques; therefore, I utilized the trial and test corpora provided by the organizers of SemEval 2012 as evaluation corpora. I did this both to expand the size of my evaluation corpus as well as to potentially examine the effects that on the efficacy of these models that different corpora, and different sizes in corpora, can have.

2.2 Corpus Cleaning

While examining the trial and test corpora, I noticed two major problems: the first with HTML entities the second with inflected vs. lemmatized versions of words. The context texts of the trial and test corpora were not always in plain text, and in particular the context texts contained HTML entities. Since some of my methods (notably the n-gram frequency counts) use the context of a target word, I decided to create a cleaner version of the corpora. For most of the HTML entities, such as dash and quote, (ex. ” “ etc.), I simply replaced each entity with the symbol that it referred to. Replacing the apostrophe HTML entity was slightly more difficult, for two reasons. First, the token containing the apostrophe HTML entity was separated from the earlier word that it would normally be attached to, and second I faced the problem of determining whether to keep the contractions as contractions and link the abbreviated token with the previous one, or to keep them separate. I decided to link them because the contractions were in the text, and it was feasible that contracts might affect the difficulty of understanding a given sentence, particularly for new English speakers. In addition to these changes, I should note that in some cases the HTML entities had a space between the first part of their tag and the semicolon denoting the end of their tag. (ex.: &#8220 ;) I accounted for this in my corpus cleaning but it is still an important thing to note for those wishing to use these corpora.

The second major problem was the question of inflection vs. Lemmatization. Namely, in some examples the target word in the sentence was in an inflected form, but the substitute candidates were in their lemmatized (or root) forms. For example:

<corpus lang="english">
<lexelt item="bright.a">
<instance id="5">
<context> In fact, during at least six distinct periods in Army history since World War I , lack of trust and confidence in senior leaders caused the so-called best and <head>brightest</head> to leave the Army in droves .</context>

With the potential substitutes being: motivated; bright; capable; clever; promising; intelligent; sharp; most able. As you can see, only one of these substitutes is in the correct form (most able), and phrases such as 'best and sharp' or 'best and 'intelligent', don't make sense in context and would likely not occur in a corpus of English N-Grams.

In addition to the conundrum that this placed on using N-Gram models, the question arose as to whether inflected the potential substitutes was a good idea at all, as the evaluators did not have inflected forms when judging and ranking the simplicity of potential substitutes. I determined that not having inflected forms would make any attempt at using higher order N-Grams useless, so I decided to inflect the potential substitutes before ranking

them. I also altered the gold standard so that they would match the words I was ranking, for the sake of making evaluation go more smoothly. To do this, I used two modules: the Nodebox:Linguistics module[1] for inflected nouns and verbs and the Pattern.en module[2] for generating superlative and comparative forms of adjectives. Using both modules, I compared the root form of the word, provided in the corpora, against the inflected form in context, and then inflected all of the potential substitutes with this same inflection. I left substitutes that could not be effectively inflected by the modules in their lemmatized forms. Also, the adverbs in the context sentence were always in their root form in these corpora, so I did not change them.

3 Methods

3.1 Word Length

Ranking the simplicity of a word by ascending word length is by far the simplest implementation of lexical simplification I utilized. This implementation simply involved collecting all of the potential substitutions for each example in a list and sorting them by the length of the word. I performed this method on both the lemmatized and inflected forms of the word, to examine whether this technique would be useful in both cases. On the whole, this method was my own baseline case, a very simple lexical simplification tool that any truly useful tool would probably have to be better than.

3.2 Word Sense Count

My second lexical simplification method utilized the WordNet database to provide sense counts for each substitute. This involved querying WordNet for all senses of a certain part of speech, given in the corpus, for each of the lemmatized word and then collecting them in multiple dictionary data structures, one for each example. Then, sorting by the value (number of senses) I generated simplicity rankings based on ascending and descending sense counts. The reason for both ascending and descending order was that I had competeting intuitions on which ordering would yield more correct simplicity rankings.

[1]http://nodebox.net/code/index.php/Linguistics

[2]http://www.clips.ua.ac.be/pages/pattern-en

3.3 Word Sense Entropy

Somewhat related to the rankings based on the number of senses, I also calculated the entropy of the word senses for each substitute. The intuition was that the distribution of frequency between word senses might affect how easy to understand a word is and thus the ranking system. In this case I also queried for WordNet for each of the word senses, but instead of simply summing their number, I queried WordNet for the frequency of each of these word senses and then computed the entropy of their occurence. I then created rankings both by ascending entropy values and descending entropy values, as I had little intuition as to which would yield better results.

3.4 Simple English Wikipedia N-Grams

Utilizing the Simple English Wikipedia for the sake of ranking based on N-gram frequency was one of the more arduous methods to implement, mainly due to data collection and pre-processing involved. The first step was to download a data dump of the Simple English Wikipedia. It is somewhat difficult to find data dumps that only include the articles and no meta-data, but the meta-data documents are of a fairly uniform format and can be deleted after download relatively easily. The next step was to extract this data, and then to convert all of the downloaded documents from a wiki-mark up format into an XML format. Finally, I converted all of these XML documents into one large text documents, as my frequency counts would be based on this corpus as a whole.[3]

Instead of generating N-gram counts from this corpus directly, I decided that I would create all of the potential N-grams I was searching for first, and then scan this corpus to accumulate data on their frequency. It is important to note here that I consider each substitution to be a 'unigram,' even when it is potentially made up of multiple words (such as 'most difficult'), and that I used the inflected forms of the substitutions in order to avoid the problems with lemmatized forms in context expressed earlier. I then ranked each substitute in descending or-

[3]This process, as well as useful tools for completing it, are outlined here (http://blog.afterthedeadline.com/2009/12/04/generating-a-plain-text-corpus-from-wikipedia/)

der of frequency, counting unigrams, then bigrams, and trigrams. For those interested in utilizing Simple English Wikipedia for their own purposes or for replicating my results, I advise that this process will take at least a day, potentially more depending on your hardware. Additionally, I used a stupid back off parameter of 1/100 for unigrams to bigrams and 1/10000 for unigrams to trigrams when calculating frequencies based on bigrams and trigrams which did not exist in the Simple English Wikipedia.

3.5 British National Corpus N-Grams

In contrast to the Simple Wikipedia Corpus, I was provided with the British National Corpus, which is constituted by one hundred million words, over 6 million lines in length, in a format suitable for N-gram frequency creation. I utilized the same general methods for creating the N-Gram frequency counts for this corpus as for the Simple Wikipedia Corpus above, although the size of the corpus was much larger and thus had a much longer processing time.

4 Results

The results for the experiments described above may be a bit confusing, as I utilized two evaluation systems and two corpora to generate answers from. In order to minimize confusion, I will divide the results table by corpus, although the two evaluation scores for each system in each corpus will be in the same table. The first evaluation score "SemEval" represents the SemEval pairwise evaluation scheme described in section 1.3 above. In contrast the 'Simplest' evaluation scheme is a simple Precision measurement of how many 'simplest' words did a system get correct. In order to minimize confusion, "SemEval" scores are given in decimal, whereas "Simplest Scores" are given as a Percent Correct.

4.1 Length, Word Sense Count, Word Sense Entropy

This first table holds the SemEval and Simplest scores for the Length, Maximum Word Sense Count, Minimum Word Sense Count, Maximum Word Sense Entropy, and Minimum Word Sense Entropy evaluated on the trial corpus. These are grouped together as they were the less complex methods implemented in this paper, and some had similar results.

Table 2 holds the results for the same systems, but evaluated on the Test corpus rather than the trial corpus.

Table 1: Simple Systems on the Trial Corpus

System	SemEval	Precision(percent)
Length	0.2059	36.88
Maximum WS	0.2175	38.87
Minimum WS	-0.2192	12.96
Maximum SE	0.2132	35.22
Minimum SE	-0.2149	15.28

The key features of these results are 1) Maximum

Table 2: Simple Systems on the Test Corpus

System	SemEval	Precision(percent)
Length	0.2757	49.50
Maximum WS	0.3037	47.87
Minimum WS	-0.3050	11.75
Maximum SE	0.2622	42.61
Minimum SE	-0.2634	12.04

Word Sense(WS), Maximum Sense Entropy(SE) and Length have fairly good scores on both the trial and test corporain both the SemEval and Precision categories 2) That Minimum WS and Minimum SE have low scores on both the trial and test corpora in both the SemEval and Precision categories, and 3) There was a larger dichotomoy between these two groups' scores in the Test Corpora than the Trial Corpora.

4.2 Simple English Wikipedia

The results for the Simple English Wikipedia N-Gram system will also be divided by trial and test corpus, and by whether unigrams, bigrams, or trigrams were utilized. This system is considered alone as it was implemented to determine whether a corpus supervised for simplicity could generate N-Gram frequencies related to simplicity more effectively.

Table 4 holds the results for the same impe-

Table 3: Simple Wikipedia (SW) N-Gram Frequency on the Trial Corpus

System	SemEval	Precision(percent)
SW Unigram	0.3853	47.84
SW Bigram	0.1733	36.54
SW Trigram	0.085	30.56

mentations, but on the test corpus.

There are a number of key points to be gleaned

Table 4: Simple Wikipedia (SW) N-Gram Frequency on the Test Corpus

System	SemEval	Precision(percent)
SW Unigram	0.4686	57.98
SW Bigram	0.2428	46.99
SW Trigram	0.1028	34.66

from these results. 1) The Simple Wikipedia Unigram system is a fairly accurate system, having a significantly higher SemEval and Simplest Precision score than any of the simple systems. 2) The Simple Wikipedia systems follow the same pattern of doing better on the test corpus than the trial corpus in both the SemEval and Simplest Precision categories. 3) While the pattern of SemEval and Precision being directed related continues, the rate of decline of the Simplest Precision score from Unigram to Bigram and Bigram to Trigram is much lower than the SemEval Precision score's decline.

4.3 British National Corpus N-Gram Frequency

The British National Corpus(BNC) N-Gram Frequency system was designed to examine specifically if more common words are simpler. (or if simpler words are more common, the system would not be able to distinguish) It's results will also be split by corpus.

Table 6 holds the results for the same system, but evaluated on the Test Corpus.

Table 5: BNC N-Gram Frequency on the Trial Corpus

System	SemEval	Precision(percent)
BNC Unigram	0.3780	44,85
BNC Bigram	0.3004	43.52
BNC Trigram	0.1745	36.21

Table 6: BNC N-Gram Frequency on the Test Corpus

System	SemEval	Precision(percent)
BNC Unigram	0.4657	59.91
BNC Bigram	0.3610	55.35
BNC Trigram	0.2142	44.94

The British National Corpus N-Gram system's results follow most of the same patterns as the Simple Wikipedia system's results. Unigram based was better than Bigram based which was better than Trigram based, the test corpus results were better than the trial corpus results, the SemEval and Simplest Precision were fairly closely related, and the 'best' system, the BNC Unigram system, had remarkably similar SemEval and Simplest Precision scores compared to the Simplie Wikipedia Unigram system. One important difference is that while the higher order N-Gram systems have lower results for the BNC based system, they are better than their equivalents in the Simple Wikipedia system.

All of the results presented above are avaiable in Appendix A at the back of this paper in a unified table.

5 Analysis

The results of the experiments described in this paper show a number of important and interesting information and trends.

5.1 Entropy, Word Senses, and Length

Two interesting patterns are that ranking based on descending maximum sense entropy and maximum sense number are better ranking schemes than ranking based on minimum sense entropy and minimum

sense number. This is not necessarily intuitive, and when I initially chose to implement these methods, I thought that minimum sense entropy and minimum sense number would be better, on the theory that since the individuals who created the gold standard for this task were new to speaking English, they would have greater problems understanding words with multiple senses as understanding the way that the word translates in the specific context could be a problem. Therefore, words with fewer senses or words which have some senses which dominate their occurrence would be easier to understand. This however is not the case, as demonstrated above, since the minimum sense entropy and count have the lowest scores out of all methods implemented. I hypothesize that the 'goodness' of maximum entropy and sense number is due to their relationship with word frequency (words with many senses and high sense entropy are likely to be common words), but at the time of writing this paper I have not been able to demonstrate this. In addition, the results above showcase that maximum sense number is better than maximum sense entropy in both the trial and test cases, although they are fairly close, suggesting that the number of senses matter more than their distribution in determining lexical simplicity.

Ranking based on word length has similar scores to maximum sense entropy and maximum sense number in both the trial an test cases, suggesting that all three factors have a similar level of influence or relationship to the simplicity of a word.

5.2 Supervised vs. Size in N-Gram Frequency

The differences between the scores based on N-Gram frequencies generated by the Simple English Wikipedia corpus and the British National Corpus, and among the scores based on each for unigrams, bigrams, and trigrams, show a number of interesting ideas. First, the Simple English Wikipedia Unigram system had better SemEval evaluation scores that the BNC systems. Considering that the Simple English Wikipedia is far less than 1/6 the size of the BNC, the supervised nature of the Simple English Wikipedia as 'simple' seems to have major value. In addition, the fact that the BNC system was so good suggests that basic word frequency is probably one of the best features for determining word simplicity. Furthermore, the bigram and trigram systems actually had worse scores for both the Simple English Wikipedia and the BNC. At the time of writing this paper, I have been unable to determine why this is the case, as having more data would normally lead to better results, not worse results. My hypothesis at the moment is that the corpora are too small to have a significant set of bigrams and trigrams for distinguishing between frequent and infrequent bigrams and trigrams. This is supported by the fact that while the bigram and trigram systems were worse for the BNC, the drop in performance between unigram-bigram and bigram-trigram was far less for the BNC, the larger corpus, than the Simple English Wikipedia. This is also supported by the fact that the decline in the Simplest Precision score was much less marked than in the SemEval score, suggesting that the bigram and trigram systems were more able to choose 'the simplest' word correctly (the one for which, definitionally, there would be the most frequency data) at a better rate than the less simple words. Another potential explanation is a bad stupid back-off constant; however, I tested the system not using any form of back-off, and my results were not appreciably different.

5.3 Simplest vs. SemEval

Another important trend to note is that the SemEval evaluation system and evaluating only the 'simplest' word compared to the 'simplest' word in the gold standard were fairly strongly linked in most cases. Systems with high SemEval scores had high 'simplest' scores, low had low, etc. However, some systems that had fairly close SemEval scores had 'simplest' scores which were reversed in terms of their size (i.e. the system with the larger SemEval score had the smaller 'simplest' score). The most important example of this occurred between the Simple English Wikipedia unigram system and the BNC unigram system, the first having the higher SemEval score and the second having the higher 'simplest' score. (See above) Overall, this suggests that if we are looking for systems for the practical purpose of converting texts, the SemEval evaluation scheme is probably not the best to use for determining which system should be utilized.

5.4 Trial vs. Test

Finally, all of my systems did better on the test data than the trial data. This is not a problem, as I did not utilize any machine learning techniques; however it does demonstrate that such techniques might not have a high degree of efficacy, as what constitutes 'simple' seems to be very different between the two corpora.

6 Conclusion

In this paper, I presented a number of systems designed for the English Lexical Simplification task of SemEval 2012. Based on the SemEval evaluation system, I obtained the best results using frequencies from the Simple English Wikipedia, with a score of 0.4686. Evaluating only based on the 'simplest' word, I obtained the best results using frequencies from the British National Corpus with a precision of 59.91%. While the results obtained by the Simple English Wikipedia were not significantly better than a Simple Frequency count (and were lower than the baseline provided by SemEval which used the Google IT Corpus), the Simple English Wikipedia corpus was SIGNIFICANTLY smaller than either of the these corpora. This would normally degrade the results of any N-Gram method, so the fact that the Simple English Wikipedia system obtained results similar to the results of much larger corpora suggests that using supervised 'simple' corpora to obtain frequency counts has significant potential.

Finally, these results suggest that any form of lexical manipulation might be benefited by using a corpus supervised for that purpose. One example that comes to mind is converting modern books to older lexicons or older books to more modern lexicons using old and modern books as frequency corpora respectively. Another examle would be converting texts for individuals new to its native language using corpora written by individuals who speak the same language as the potential reader and picked up the texts' native language as a second or third language.

7 Future Work

This project has presented a number of interesting avenues of research and future work. My first step in continuing this project would be to fine-tune my pre-processing methods, in particular the inflection methods. While these methods were fairly good, there were a few notable cases where a word would be inflected incorrectly, such as 'most strenuous' not being counted as a superlative and being inflected to be 'most most strenuous.' This did not happen often, but I feel that handling these cases would benefit the systems described above.

I would also be very interested in testing these systems on corpora tagged by different audiences (some mentioned in the Related Works section) such as children and aphasic readers, or applying these principles to other languages and seeing if they are language specific.

Finally, I would like to see if utilizing somewhat supervised corpora could work on other lexical manipulation tasks, as mentioned in the conclusion.

References

Sandra M. Aluisio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards brazillian portuguese authomatic text simplification systems. In *Proceedings of the eigth ACM symposium on Document engineering*.

Or Biran, Samuel Brody, and Noemie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

John Carroll, Yvonne Canning, Guido Minnen, Siobhan Devlin, Darren Pearce, and John Tait. 1998-1999. Simplifying text for language-imparied readers. In *Proceedings of the 9th Conferences of the European Chapter of the ACL*.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems.2010*.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*.

Lucia Specia, Sujay K. Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation*.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Proceedings of the NAACL 2010*.