SemEval-2012 Task 1: English Lexical Simplification

Lucia Specia
Department of Computer Science
University of Sheffield
L.Specia@sheffield.ac.uk

Sujay Kumar Jauhar
Research Group in Computational Linguistics
University of Wolverhampton
Sujay.KumarJauhar@wlv.ac.uk

Rada Mihalcea
Department of Computer Science and Engineering
University of North Texas
rada@cs.unt.edu

Abstract

We describe the English Lexical Simplification task at SemEval-2012. This is the first time such a shared task has been organized and its goal is to provide a framework for the evaluation of systems for lexical simplification and foster research on context-aware lexical simplification approaches. The task requires that annotators and systems rank a number of alternative substitutes – all deemed adequate – for a target word in context, according to how "simple" these substitutes are. The notion of simplicity is biased towards non-native speakers of English. Out of nine participating systems, the best scoring ones combine context-dependent and context-independent information, with the strongest individual contribution given by the frequency of the substitute regardless of its context.

1 Introduction

Lexical Simplification is a subtask of Text Simplification (Siddharthan, 2006) concerned with replacing words or short phrases by simpler variants in a context aware fashion (generally synonyms), which can be understood by a wider range of readers. It generally envisages a certain human target audience that may find it difficult or impossible to understand complex words or phrases, e.g., children, people with poor literacy levels or cognitive disabilities, or second language learners. It is similar in many respects to the task of Lexical Substitution (McCarthy and Navigli, 2007) in that it involves determining adequate substitutes in context, but in this case on the basis of a predefined criterion: simplicity.

A common pipeline for a Lexical Simplification system includes at least three major components: (i) complexity analysis: selection of words or phrases in a text that are considered complex for the reader and/or task at hand; (ii) substitute lookup: search for adequate replacement words or phrases deemed complex in context, e.g., taking synonyms (with the same sense) from a thesaurus or finding similar words/phrases in a corpus using distributional similarity metrics; and (iii) context-based ranking: ranking of substitutes according to how simple they are to the reader/task at hand.

As an example take the sentence: *"Hitler committed terrible atrocities during the second World War."* The system would first identify complex words, e.g. *atrocities*, then search for substitutes that might adequately replace it. A thesaurus lookup would yield the following synonyms: *abomination*, *cruelty*, *enormity* and *violation*, but *enormity* should be dropped as it does not fit the context appropriately. Finally, the system would determine the simplest of these substitutes, e.g., *cruelty*, and use it to replace the complex word, yielding the sentence: *"Hitler committed terrible cruelties during the second World War."*.

Different from other subtasks of Text Simplification like Syntactic Simplification, which have been relatively well studied, Lexical Simplification has received less attention. Although a few recent attempts explicitly address dependency on context (de Belder et al., 2010; Yatskar et al., 2010; Biran et al., 2011; Specia, 2010), most approaches are context-independent (Candido et al., 2009; Devlin and Tait, 1998). In addition, a general deeper understanding

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 347–355,
Montréal, Canada, June 7-8, 2012.

of the problem is yet to be gained. As a first attempt to address this problem in the shape of a shared task, the English Simplification task at SemEval-2012 focuses on the third component, which we believe is the core of the Lexical Simplification problem.

The SemEval-2012 shared task on English Lexical Simplification has been conceived with the following main purposes: advancing the state-of-the-art Lexical Simplification approaches, and providing a common framework for evaluation of Lexical Simplification systems for participants and other researchers interested in the field. Another central motive of such a shared task is to bring awareness to the general vagueness associated with the notion of lexical simplicity. Our hypothesis is that in addition to the notion of a target application/reader, the notion of simplicity is highly context-dependent. In other words, given the same list of substitutes for a given target word with the **same sense**, we expect different orderings of these substitutes in different contexts. We hope that participation in this shared task will help discover some underlying traits of lexical simplicity and furthermore shed some light on how this may be leveraged in future work.

2 Task definition

Given a short context, a target word in English, and several substitutes for the target word that are deemed adequate for that context, the goal of the English Simplification task at SemEval-2012 is to rank these substitutes according to how "simple" they are, allowing ties. Simple words/phrases are loosely defined as those which can be understood by a wide range of people, including those with low literacy levels or some cognitive disability, children, and non-native speakers of English. In particular, the data provided as part of the task is annotated by **fluent but non-native speakers of English**.

The task thus essentially involves comparing words or phrases and determining their order of complexity. By ranking the candidates, as opposed to categorizing them into specific labels (simple, moderate, complex, etc.), we avoid the need for a fixed number of categories and for more subjective judgments. Also ranking enables a more natural and intuitive way for humans (and systems) to perform annotations by preventing them from treating each individual case in isolation, as opposed to relative to each other. However, the inherent subjectivity introduced by ranking entails higher disagreement among human annotators, and more complexity for systems to tackle.

3 Corpus compilation

The trial and test corpora were created from the corpus of SemEval-2007 shared task on Lexical Substitution (McCarthy and Navigli, 2007). This decision was motivated by the similarity between the two tasks. Moreover the existing corpus provided an adequate solution given time and cost constraints for our corpus creation. Given existing contexts with the original target word replaced by a placeholder and the lists of substitutes (including the target word), annotators (and systems) are required to rank substitutes in order of simplicity for each context.

3.1 SemEval-2007 - LS corpus

The corpus from the shared task on Lexical Substitution (LS) at SemEval-2007 is a selection of sentences, or *contexts*, extracted from the English Internet Corpus of English (Sharoff, 2006). It contains samples of English texts crawled from the web.

This selection makes up the dataset of a total of $2,010$ contexts which are divided into **Trial** and **Test** sets, consisting of 300 and 1710 contexts respectively. It covers a total of 201 (mostly polysemous) target words, including nouns, verbs, adjectives and adverbs, and each of the target words is shown in 10 different contexts. Annotators had been asked to suggest up to three different substitutes (words or short phrases) for each of the target words within their contexts. The substitutes were lemmatized unless it was deemed that the lemmatization would alter the meaning of the substitute. Annotators were all native English speakers and each annotated the entire dataset. Here is an example of a context for the target word "bright":

```
<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> ... </lexelt>

The gold-standard document contains each target word along with a ranked list of its possible substitutes, e.g., for the context above, three annotators suggested "intelligent" and "clever" as substitutes for "bright", while only one annotator came up with "smart":

bright.a 1:: intelligent 3; clever 3; smart 1;

3.2 SemEval-2012 Lexical Simplification corpus

Given the list of contexts and each respective list of substitutes we asked annotators to rank substitutes for each individual context in ascending order of complexity. Since the notion of textual simplicity varies from individual to individual, we carefully chose a group of annotators in an attempt to capture as much of a common notion of simplicity as possible. For practical reasons, we selected annotators with high proficiency levels in English as second language learners - all with a university first degree in different subjects.

The Trial dataset was annotated by four people while the Test dataset was annotated by five people. In both cases each annotator tagged the complete dataset.

Inter-annotator agreement was computed using an adaptation of the **kappa** index with pairwise rank comparisons (Callison-Burch et al., 2011). This is also the primary evaluation metric for participating systems in the shared task, and it is covered in more detail in Section 4.

The inter-annotator agreement was computed for each pair of annotators and averaged over all possible pairs for a final agreement score. On the Trial dataset, a kappa index of 0.386 was found, while for the Test dataset, a kappa index of 0.398 was found. It may be noted that certain annotators disagreed considerably with all others. For example, on the Test set, if annotations from one judge are removed, the average inter-annotator agreement rises to 0.443. While these scores are apparently low, the highly subjective nature of the annotation task must be taken into account. According to the reference values for other tasks, this level of agreement is considered "moderate" (Callison-Burch et al., 2011).

It is interesting to note that higher inter-annotator agreement scores were achieved between annotators with similar language and/or educational backgrounds. The highest of any pairwise annotator agreement (0.52) was achieved between annotators of identical language and educational background, as well as very similar levels of English proficiency. High agreement scores were also achieved between annotators with first languages belonging to the same language family.

Finally, it is also worth noticing that this agreement metric is highly sensitive to small differences in annotation, thus leading to overly pessimistic scores. A brief analysis reveals that annotators often agree on clusters of simplicity and the source of the disagreement comes from the rankings within these clusters.

Finally, the gold-standard annotations for the Trial and Test datasets – against which systems are to be evaluated – were generated by averaging the annotations from all annotators. This was done context by context where each substitution was attributed a score based upon the average of the rankings it was ascribed. The substitutions were then sorted in ascending order of scores, i.e., lowest score (highest average ranking) first. Tied scores were grouped together to form a single rank. For example, assume that for a certain context, four annotators provided rankings as given below, where multiple candidates between {} indicate ties:

Annotator 1: {clear} {light} {bright} {luminous} {well-lit}

Annotator 2: {well-lit} {clear} {light} {bright} {luminous}

Annotator 3: {clear} {bright} {light} {luminous} {well-lit}

Annotator 4: {bright} {well-lit} {luminous} {clear} {light}

Thus the word "clear", having been ranked 1st, 2nd, 1st and 4th by each of the annotators respectively is given an averaged ranking score of 2. Similarly "light" = 3.25, "bright" = 2.5, "luminous" = 4 and "well-lit" = 3.25. Consequently the gold-standard ranking for this context is:

Gold: {clear} {bright} {light, well-lit} {luminous}

3.3 Context-dependency

As mentioned in Section 1, one of our hypotheses was that the notion of simplicity is context-dependent. In other words, that the ordering of substitutes for different occurrences of a target word with a given sense is highly dependent on the contexts in which such a target word appears. In order to verify this hypothesis quantitatively, we further analyzed the gold-standard annotations of the Trial and Test datasets. We assume that identical lists of substitutes for different occurrences of a given target word ensure that such a target word has the same sense in all these occurrences. For every target word, we then generate all pairs of contexts containing the exact same initial list of substitutes and check the proportion of these contexts for which human annotators ranked the substitutes differently. We also check for cases where only the top-ranked substitute is different. The numbers obtained are shown in Table 1.

	Trial	Test
1) # context pairs	1350	7695
2) # 1) with same list	60	242
3) # 2) with different rankings	24	139
4) # 2) with different top substitute	19	38

Table 1: Analysis on the context-dependency of the notion of simplicity.

Although the proportion of pairs of contexts with the same list of substitutes is very low (less than 5%), it is likely that there are many other occurrences of a target word with the same sense and slightly different lists of substitutes. Further manual inspection is necessary to determine the actual numbers. Nevertheless, from the observed sample it is possible to conclude that humans will, in fact, rank the same set of words (with the same sense) differently depending on the context (on an average in 40-57% of the instances).

4 Evaluation metric

No standard metric has yet been defined for evaluating Lexical Simplification systems. Evaluating such systems is a challenging problem due to the aforementioned subjectivity of the task. Since this is a ranking task, rank correlation metrics are desirable. However, metrics such as Spearman's Rank Correlation are not reliable on the limited number of data points available for comparison on each ranking (note that the nature of the problem enforces a context-by-context ranking, as opposed to a global score), Other metrics for localized, pairwise rank correlation, such as Kendall's Tau, disregard ties, – which are important for our purposes – and are thus not suitable.

The main evaluation metric proposed for this shared task is in fact a measure of inter-annotator agreement, which is used for both contrasting two human annotators (Section 3.2) and contrasting a system output to the average of human annotations that together forms the gold-standard.

Out metric is based on the kappa index (Cohen, 1960) which in spite of many criticisms is widely used for its simplicity and adaptability for different applications. The generalized form of the kappa index is

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ denotes the proportion of times two annotators agree and $P(E)$ gives the probability of agreement by chance between them.

In order to apply the kappa index for a ranking task, we follow the method proposed by (Callison-Burch et al., 2011) for measuring agreement over judgments of translation quality. This method defines $P(A)$ and $P(E)$ in such a way that it now counts agreement whenever annotators concur upon the order of pairwise ranks. Thus, if one annotator ranked two given words 1 and 3, and the second annotator ranked them 3 and 7 respectively, they are still in agreement. Formally, assume that two annotators $A1$ and $A2$ rank two instance a and b. Then $P(A) =$ the proportion of times $A1$ and $A2$ agree on a ranking, where an occurrence of agreement is counted whenever $rank(a < b)$ or $rank(a = b)$ or $rank(a > b)$.

$P(E)$ (the likelihood that annotators $A1$ and $A2$ agree by chance) is based upon the probability that both of them assign the same ranking order to a and b. Given that the probability of getting $rank(a < b)$ by any annotator is $P(a < b)$, the probability that *both* annotators get $rank(a < b)$ is $P(a < b)^2$ (agreement is achieved when $A1$ assigns $a < b$ by chance and $A2$ also assigns $a < b$). Similarly, the

probability of chance agreement for $rank(a = b)$ and $rank(a > b)$ are $P(a = b)^2$ and $P(a > b)^2$ respectively. Thus:

$$P(E) = P(a < b)^2 + P(a = b)^2 + P(a > b)^2$$

However, the counts of $rank(a < b)$ and $rank(a > b)$ are inextricably linked, since for any particular case of $a_1 < b_1$, it follows that $b_1 > a_1$, and thus the two counts must be incremented equally. Therefore, over the entire space of ranked pairs, the probabilities remain exactly the same. In essence, after counting for $P(a = b)$, the remaining probability mass is equally split between $P(a < b)$ and $P(a > b)$. Therefore:

$$P(a < b) = P(a > b) = \frac{1 - P(a = b)}{2}$$

Kappa is calculated for every pair of ranked items for a given context, and then averaged to get an overall kappa score:

$$\kappa = \frac{\sum_{n=1}^{|N|} \frac{P_n(A) - P_n(E)}{1 - P_n(E)}}{|N|}$$

where N is the total number of contexts, and $P_n(A)$ and $P_n(E)$ are calculated based on counts extracted from the data on the particular context n.

The functioning of this evaluation metric is illustrated by the following example:

> **Context:** During the siege, George Robertson had appointed Shuja-ul-Mulk, who was a _____ boy only 12 years old and the youngest surviving son of Aman-ul-Mulk, as the ruler of Chitral.
>
> **Gold:** {intelligent} {clever} {smart} {bright}
>
> **System:** {intelligent} {bright} {clever, smart}

Out of the 6 distinct unordered pairs of lexical items, *system* and *gold* agreed 3 times. Consequently, $P_n(A) = \frac{3}{6}$. In addition, $count(a = b) = 1$. Thus, $P_n(a = b) = \frac{1}{12}$. Which gives a $P(E) = \frac{41}{96}$ and the final kappa score for this particular context of 0.13.

The statistical significance of the results from two systems A and B is measured using the method of **Approximate Randomization**, which has been shown to be a robust approach for several NLP tasks (Noreen, 1989). The randomization is run $1,000$ times and if the p-value is ≤ 0.05 the difference between systems A and B is asserted as being statistically significance.

5 Baselines

We defined three baseline lexical simplification systems for this task, as follows.

L-Sub Gold: This baseline uses the gold-standard annotations from the Lexical Substitution corpus of SemEval-2007 as is. In other words, the ranking is based on the *goodness of fit* of substitutes for a context, as judged by human annotators. This method also serves to show that the Lexical Substitution and Lexical Simplification tasks are indeed different.

Random: This baseline provides a randomized order of the substitutes for every context. The process of randomization is such that is allows the occurrence of ties.

Simple Freq.: This simple frequency baseline uses the frequency of the substitutes as extracted from the Google Web 1T Corpus (Brants and Franz, 2006) to rank candidate substitutes within each context.

The results in Table 2 show that the "L-Sub Gold" and "Random" baselines perform very poorly on both Trial and Test sets. In particular, the reason for the poor scores for "L-Sub Gold" can be attributed to the fact that it yields many ties, whereas the gold-standard presents almost no ties. Our kappa metric tends to penalize system outputs with too many ties, since the probability of agreement by chance is primarily computed on the basis of the number of ties present in the two rankings being compared (see Section 4).

The "Simple Freq." baseline, on the other hand, performs very strongly, in spite of its simplistic approach, which is entirely agnostic to context. In fact it surpasses the average inter-annotator agreement on both Trial and Test datasets. Indeed, the scores on the Test set approach the best inter-annotator agreement scores between any two annotators.

	Trial	Test
L-Sub Gold	0.050	0.106
Random	0.016	0.012
Simple Freq.	0.397	0.471

Table 2: Baseline kappa scores on trial and test sets

6 Results and Discussion

6.1 Participants

Five sites submitted one or more systems to the task, totaling nine systems:

ANNLOR-lmbing: This system (Ligozat et al., 2012) relies on language models probabilities, and builds on the principle of the Simple Frequency baseline. While the baseline uses Google n-grams to rank substitutes, this approach uses Microsoft Web n-grams in the same way. Additionally characteristics, such as the contexts of each term to be substituted, were integrated into the system. Microsoft Web N-gram Service was used to obtain log likelihood probabilities for text units, composed of the lexical item and 4 words to the left and right from the surrounding context.

ANNLOR-simple: The system (Ligozat et al., 2012) is based on Simple English Wikipedia frequencies, with the motivation that the language used in this version of Wikipedia is targeted towards people who are not first-language English speakers. Word n-grams (n = 1-3) and their frequencies were extracted from this corpus using the Text-NSP Perl module and a ranking of the possible substitutes of a target word according to these frequencies in descending order was produced.

EMNLPCPH-ORD1: The system performs a series of pairwise comparisons between candidates. A binary classifier is learned purpose using the Trial dataset and artificial unlabeled data extracted based on Wordnet and a corpus in a semi-supervised fashion. A co-training procedure that lets each classifier increase the other classifier's training set with selected instances from the unlabeled dataset is used. The features include word and character n-gram probabilities of candidates and contexts using web corpora, distributional differences of candidate in a corpus of "easy" sentences and a corpus of normal sentences, syntactic complexity of documents that are similar to the given context, candidate length, and letter-wise recognizability of candidate as measured by a trigram LM. The first feature sets for co-training combines the syntactic complexity, character trigram LM and basic word length features, resulting in 29 features against the remaining 21.

EMNLPCPH-ORD2: This is a variant of the EMNLPCPH-ORD1 system where the first feature set pools all syntactic complexity features and Wikipedia-based features (28 features) against all the remaining 22 features in the second group.

SB-mmSystem: The approach (Amoia and Romanelli, 2012) builds on the baseline definition of simplicity using word frequencies but attempt at defining a more linguistically motivated notion of simplicity based on lexical semantics considerations. It adopts different strategies depending on the syntactic complexity of the substitute. For one-word substitutes or common collocations, the system uses its frequency from Wordnet as a metric. In the case of multi-words substitutes the system uses "relevance" rules that apply (de)compositional semantic criteria and attempts to identify a unique content word in the substitute that might better approximate the whole expression. The expression is then assigned the frequency associated to this content word for the ranking. After POS tagging and sense disambiguating all substitutes, hand-written rules are used to decompose the meaning of a complex phrase and identify the most relevant word conveying the semantics of the whole.

UNT-SimpRank: The system (Sinha, 2012) uses external resources, including the Simple English Wikipedia corpus, a set of Spoken English dialogues, transcribed into machine readable form, WordNet, and unigram frequencies (Google Web1T data). SimpRank scores each substitute by a sum of its unigram frequency, its

frequency in the Simple English Wikipedia, its frequency in the spoken corpus, the inverse of its length, and the number of senses the substitute has in WordNet. For a given context, the substitutes are then reverse-ranked based on their simplicity scores.

UNT-SimpRankLight: This is a variant of SimpRank which does not use unigram frequencies. The goal of this system is to check whether a memory and time-intensive and non-free resource such as the Web1T corpus makes a difference over other free and lightweight resources.

UNT-SaLSA: The only resource SaLSA depends on is the Web1T data, and in particular only 3-grams from this corpus. It leverages the context provided with the dataset by replacing the target placeholder one by one with each of the substitutes and their inflections thus building sets of 3-grams for each substitute in a given instance. The score of any substitute is then the sum of the 3-gram frequencies of all the generated 3-grams for that substitute.

UOW-SHEF-SimpLex: The system (Jauhar and Specia, 2012) uses a linear weighted ranking function composed of three features to produce a ranking. These include a context sensitive n-gram frequency model, a bag-of-words model and a feature composed of simplicity oriented psycholinguistic features. These three features are combined using an SVM ranker that is trained and tuned on the Trial dataset.

6.2 Pairwise kappa

The official task results and the ranking of the systems are shown in Table 3.

Firstly, it is worthwhile to note that all the top ranking systems include features that use frequency as a surrogate measure for lexical simplicity. This indicates a very high correlation between distributional frequency of a given word and its perceived complexity level. Additionally, the top two systems involve context-dependent and context-independent features, thus supporting our hypothesis of the composite nature of the lexical simplification problem.

Rank	Team - System	Kappa
1	UOW-SHEF-SimpLex	0.496
2	UNT-SimpRank	0.471
	Baseline-Simple Freq.	0.471
	ANNLOR-simple	0.465
3	UNT-SimpRankL	0.449
4	EMNLPCPH-ORD1	0.405
5	EMNLPCPH-ORD2	0.393
6	SB-mmSystem	0.289
7	ANNLOR-lmbing	0.199
8	Baseline-L-Sub Gold	0.106
9	Baseline-Random	0.013
10	UNT-SaLSA	-0.082

Table 3: Official results and ranking according to the pairwise kappa metric. Systems are ranked together when the difference in their kappa score is not statistically significant.

Few of the systems opted to use some form of supervised learning for the task, due to the limited number of training examples given. As pointed out by some participants who checked learning curves for their systems, the performance is likely to improve with larger training sets. Without enough training data, context agnostic approaches such as the "Simple Freq." baseline become very hard to beat.

We speculate that the reason why the effects of context-aware approaches are somewhat mitigated is because of the isolated setup of the shared task. In practice, humans produce language at an even level of complexity, i.e. consistently simple, or consistently complex. In the shared task's setup, systems are expected to simplify a single target word in a context, ignoring the possibility that sometimes simple words may not be contextually associated with complex surrounding words. This not only explains why context-aware approaches are less successful than was originally expected, but also gives a reason for the good performance of context-agnostic systems.

6.3 Recall and top-rank

As previously noted, the primary evaluation metric is very susceptible to penalize slight changes, making it overly pessimistic about systems' performance. Hence, while it may be an efficient way to compare and rank systems within the framework of

a shared task, it may be unnecessarily devaluing the practical viability of approaches. We performed two post hoc evaluations that assess system output from a practical point of view. We check how well the top-ranked substitute, i.e., the simplest substitute according to a given system (which is most likely to be used in a real simplification task) compares to the top-ranked candidate from the gold standard. This is reported in the TRnk column of Table 4: the percentage of contexts in which the intersection between the simplest substitute set from a system's output and the gold standard contained *at least* one element. We note that while ties are virtually inexistent in the gold standard data, ties in the system output can affect this metric: a system that naively predicts all substitutes as the simplest (i.e., a single tie including all candidates) will score 100% in this metric.

We also measured the "recall-at-n" values for $1 \leq n \leq 3$, which gives the ratio of candidates from the top *n* substitute sets to those from the gold-standard. For a given *n*, we only consider contexts that have at least *n*+1 candidates in the gold-standard (so that there is some ranking to be done). Table 4 shows the results of this additional analysis.

Team - System	TRnk	*n*=1	*n*=2	*n*=3
UOW-SHEF-SimpLex	0.602	0.575	0.689	0.769
UNT-SimpRank	0.585	0.559	0.681	0.760
Baseline-Simple Freq.	0.585	0.559	0.681	0.760
ANNLOR-simple	0.564	0.538	0.674	0.768
UNT-SimpRankL	0.567	0.541	0.674	0.753
EMNLPCPH-ORD1	0.539	0.513	0.645	0.727
EMNLPCPH-ORD2	0.530	0.503	0.637	0.722
SB-mmSystem	0.477	0.452	0.632	0.748
ANNLOR-lmbing	0.336	0.316	0.494	0.647
Baseline-L-Sub Gold	0.454	0.427	0.667	0.959
Baseline-Random	0.340	0.321	0.612	0.825
UNT-SaLSA	0.146	0.137	0.364	0.532

Table 4: Additional results according to the top-rank (TRnk) and recall-at-*n* metrics.

These evaluation metrics favour systems that produce many ties. Consequently the baselines "L-Sub Gold" and "Random" yield overly high scores for recall-at-n for *n*=2 and *n*= 3. Nevertheless the rest of the results are by and large consistent with the rankings from the kappa metric.

The results for recall-at-2, e.g., show that most systems, on average 70% of the time, are able to find the simplest 2 substitute sets that correspond to the gold standard. This indicates that most approaches are reasonably good at distinguishing very simple substitutes from very complex ones, and that the top few substitutes will most often produce effective simplifications.

These results correspond to our experience from the comparison of human annotators, who are easily able to form clusters of simplicity with high agreement, but who strongly disagree (based on personal biases towards perceptions of lexical simplicity) on the internal rankings of these clusters.

7 Conclusions

We have presented the organization and findings of the first English Lexical Simplification shared task. This was a first attempt at garnering interest in the NLP community for research focused on the lexical aspects of Text Simplification.

Our analysis has shown that there is a very strong relation between distributional frequency of words and their perceived simplicity. The best systems on the shared task were those that relied on this association, and integrated both context-dependent and context-independent features. Further analysis revealed that while context-dependent features are important in principle, their applied efficacy is somewhat lessened due to the setup of the shared task, which treats simplification as an isolated problem.

Future work would involve evaluating the importance of context for lexical simplification in the scope of a simultaneous simplification to all the words in a context. In addition, the annotation of the gold-standard datasets could be re-done taking into consideration some of the features that are now known to have clearly influenced the large variance observed in the rankings of different annotators, such as their background language and the education level. One option would be to select annotators that conform a specific instantiation of these features. This should result in a higher inter-annotator agreement and hence a simpler task for simplification systems.

Acknowledgments

We would like to thank the annotators for their hard work in delivering the corpus on time.

References

Marilisa Amoia and Massimo Romanelli. 2012. SBmmSystem: Using Decompositional Semantics for Lexical Simplification. In *English Lexical Simplification. Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada.

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*, pages 496–501, Portland, Oregon.

Thorsten Brants and Alex Franz. 2006. The google web 1t 5-gram corpus version 1.1.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Arnaldo Candido, Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Boulder, Colorado.

J Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, April.

Jan de Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of Itec2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*, Kortrijk, Belgium.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Sujay Kumar Jauhar and Lucia Specia. 2012. UOW-SHEF: SimpLex - Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. In *English Lexical Simplification. Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada.

Anne-Laure Ligozat, Cyril Grouin, Anne Garcia-Fernandez, and Delphine Bernhard. 2012. ANNLOR: A Naive Notation-system for Lexical Outputs Ranking. In *English Lexical Simplification. Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada.

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), Prague, Czech Republic*, pages 48–53.

E. Noreen. 1989. Computer-intensive methods for testing hypotheses. New York: Wiley.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Ravi Sinha. 2012. UNT-SimpRank: Systems for Lexical Simplification Ranking. In *English Lexical Simplification. Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, Montreal, Canada.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, PROPOR'10, pages 30–39, Berlin, Heidelberg. Springer-Verlag.

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 *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368, Los Angeles, California.