UNT-SIMPRANK: Systems for Lexical Simplification Ranking

Ravi Sinha
University of North Texas
1155 Union Circle #311277
Denton, Texas
76203-5017
RaviSinha@my.unt.edu

Abstract

This paper presents three systems that took part in the lexical simplification task at SEMEVAL 2012. Speculating on what the concept of simplicity might mean for a word, the systems apply different approaches to rank the given candidate lists. One of the systems performs second-best (statistically significant) and another one performs third-best out of 9 systems and 3 baselines. Notably, the third-best system is very close to the second-best, and at the same time much more resource-light in comparison.

1 Introduction

Lexical simplification (described in (Specia et al., 2012)) is a newer problem that has arisen following a recent surge in interest in the related task of lexical substitution (McCarthy et al., 2007). While lexical substitution aims at making systems generate suitable paraphrases for a target word in an instance, which do not necessarily have to be simpler versions of the original, it has been speculated that one possible use of the task could be lexical simplification, in particular in the realm of making educational text more readable for non-native speakers.

The task of lexical simplification, which thus derives from lexical substitution, uses the same data set, and has been introduced at the 6th International Workshop on Semantic Evaluation (SEMEVAL 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012). Instead of asking systems to provide substitutes, the task provides the systems with all substitutes and asks them to be ranked.

The task provides several instances of triplets of a context C, a target word T, and a set of gold standard substitutes S. The systems are supposed to rank the substitutes $s_i \in S$ from the simplest to the most difficult, and match their predictions against the provided human annotations. The organizers define *simple* loosely as words that can be understood by a wide variety of people, regardless of their literacy and cognitive levels, age, and regional backgrounds.

The task is novel in that so far most work has been done on syntactic simplification and not on lexical simplification. Carroll et. al. (Carroll et al., 1998) seem to have pioneered some methodology and evaluation metrics in this field. Yatskar et. al. (Yatskar et al., 2010) use an unsupervised learning method and metadata from the Simple English Wikipedia.

2 Data

The data (trial and test, no training) have been adopted from the original lexical substitution task (McCarthy et al., 2007). The trial set has 300 examples, each with a context, a target word, and a set of substitutions. The test set has 1710 examples. The organizers provide a scorer for the task, the trial gold standard rankings, and three baselines. The data is provided in XML format, with tags identifying the lemmas, parts of speech, instances, contexts and head words. The substitutions and gold rankings are in plain text format.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 493–496,
Montréal, Canada, June 7-8, 2012.

3 Resources

Intuitively, a simple word is likely to have a high frequency in a resource that is supposed to contain simple words. Other factors that could intuitively influence simplicity would be the frequency in spoken conversation, and whether the word is polysemous or not. As such, the following resources have been selected to contribute to the metric used in ranking the substitutes.

3.1 Simple English Wikipedia

Simple English Wikipedia has been used before in simplicity analysis, as described in (Yatskar et al., 2010). It is a publicly available, smaller Wikipedia (298MB decompressed), which claims to only consist of words that are somehow *simple*. For all the substitute candidates, I count their frequencies of occurrence in this resource, and these counts serve as a factor in computing the corresponding simplicity scores (refer to Equation 1.)

3.2 Transcribed Spoken English Corpus

A set of spoken dialogues is also utilized in this project to measure simplicity. Spoken language intuitively contains more conversational words, and has the same kind of resolution power as the Simple English Wikipedia when it comes to the relative simplicity of a word. Frequency counts of all the substitute candidates in a set of dialogue corpora is computed, and used as another factor in the Equations 1 and 3.

3.3 WordNet

WordNet, as described in (Fellbaum, 1998), is a lexical knowledge base that combines the properties of a thesaurus with that of a semantic network. The basic entry in WordNet is a *synset*, which is defined as a set of synonyms. I use WordNet 3.0, which has over 150,000 unique words, over 110,000 *synsets*, and over 200,000 word-sense pairs. For each substitute, I extract the raw number of senses (for all parts of speech possible) for that word present in WordNet. This count serves as yet another factor in the proposed simplicity measure, under the hypothesis that a simple word is used very frequently, and is therefore polysemous.

3.4 Web1T Google N-gram Corpus

The Google Web 1T corpus (Brants and Franz, 2006) is a collection of English N-grams, ranging from one to five N-grams, and their respective frequency counts observed on the Web. The corpus was generated from approximately 1 trillion tokens of words from the Web, predominantly English. This corpus is also used in both SIMPRANK and SALSA systems, with the intuition that simpler words will have higher counts on the Web taken as a whole.

3.5 SaLSA

SALSA (Stand-alone Lexical Substitution Analyzer) is an in-house application which accepts as inputs sentences with target words marked distinctly, and then builds all possible 3-grams by substituting the target word with synonyms (and inflections thereof). It then queries the Web1T corpus using an in-house quick lookup application and gathers the counts for all 3-grams. Finally, it sums the counts, and assigns the aggregated scores to each corresponding synonym and outputs a reverse-ranked list of the synonyms. More detail about this methodology can be found in (Sinha and Mihalcea, 2009). SALSA uses the exact same methodology described in the paper, except that it is a stand-alone tool.

4 Experimental Setup

Figure 1 shows the general higher-level picture of how the experiments have been performed. SIMPRANK uses five resources, including the unigram frequency data, while SIMPRANKLIGHT does not use the unigram frequencies.

I hypothesize that the simplicity of a word could be represented as the Equation 1 (here $c_{word}()$ represents the frequency count of the word in a given resource).

$$\begin{aligned} simplicity(word) = \\ \frac{1}{len(word)} + c_{word}(SimpleWiki) \\ + c_{word}(Discourse) + c_{word}(WordNet) \\ + c_{word}(Unigrams) \end{aligned} \quad (1)$$

This formula is very empirical in nature, in that it has been found based on extensive experimentation

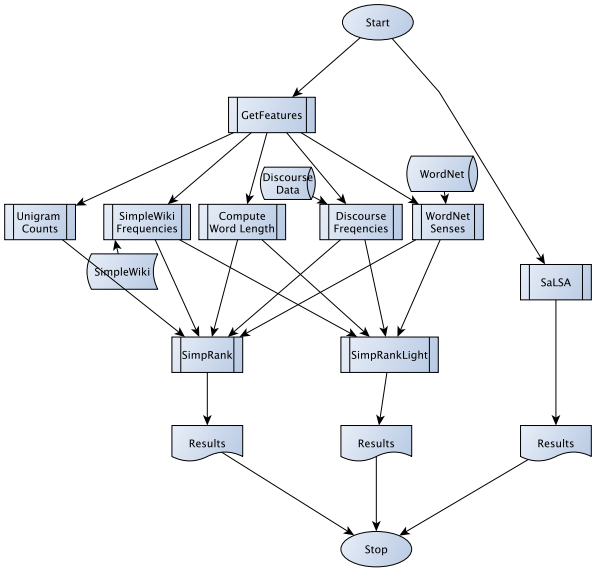

Figure 1: High-level schematic diagram of the experiments

(Table 1). It intuitively makes sense that a simple word is supposed to have high frequency counts in lexical resources that are meant to be simple by design. Formally,

$$\begin{aligned} &simplicity(word) \\ &\quad \propto frequency(SimpleResource) \\ &\quad \propto \frac{1}{length} \end{aligned} \tag{2}$$

Here, SimpleResource could be any resource that contains simple words. Apart from frequency counts, we could possibly also leverage morphology for finding simplicity. Intuitively, a 3-letter word or a 4-letter word would most likely be simpler than a word that has a longer length. This accounts for the length factor in the equations.

As Table 1 depicts, a lot of experiments were performed where the components (counts) were multiplied instead of being added, normalized instead of adding without normalization[1], and also experiments where subsets of the resources were selected. The scores obtained using the gold standard and the trial data are also shown in the table. The best combination found (experiment 8 in the table) is outlined in Equation 1.

Note however, that the Google Web1T corpus is expensive in terms of money, computation time and storage space. Thus, another set of experiments was performed (listed as experiments 1a in Table 1 leaving the unigram counts out, and it was found to work almost just as well. This system has been labeled SIMPRANKLIGHT and uses the formula in Equation 3.

$$\begin{aligned} &simplicity(word) = \\ &\quad \frac{1}{len(word)} + c_{word}(SimpleWiki) \\ &\quad + c_{word}(Discourse) + c_{word}(WordNet) \end{aligned} \tag{3}$$

The substitutes can then be sorted in the decreasing order of simplicity scores. The substitute with the highest simplicity score is hypothesized to be the simplest.

Table 1: Variants of the experiments performed

SN	System components	Method	Remarks	Score
	baseline no-change			0.05
	baseline random			0.01
	baseline unigram count (Web1T)			0.39
1	len, simplewiki, discourse, wordnet	add	normalize	0.20
1a	len, simplewiki, discourse, wordnet	add	don't normalize	**0.37**
2	len, simplewiki, discourse, wordnet	add	normalize, inc sort	-0.20
3	len, simplewiki, discourse, wordnet	multiply	don't normalize	0.25
4	simplewiki, discourse, wordnet	add	don't normalize	0.36
4a	simplewiki, discourse, wordnet	add	normalize	0.22
4b	simplewiki, discourse, wordnet	multiply	don't normalize	0.26
5	len, simplewiki, wordnet	add	don't normalize	0.36
5a	len, simplewiki, wordnet	add	normalize	0.19
5b	len, simplewiki, wordnet	multiply	don't normalize	0.26
6	len, discourse, wordnet	add	don't normalize	0.31
6a	len, discourse, wordnet	add	normalize	0.20
6b	len, discourse, wordnet	multiply	don't normalize	0.25
7	len, simplewiki, discourse	add	don't normalize	0.37
7a	len, simplewiki, discourse	add	normalize	0.22
7b	len, simplewiki, discourse	multiply	don't normalize	0.32
8	len, simplewiki, discourse, wordnet, unigrams	add	don't normalize	**0.39**
8a	len, simplewiki, discourse, wordnet, unigrams	add	normalize	0.22
8b	len, simplewiki, discourse, wordnet, unigrams	multiply	don't normalize	0.26
9	SaLSA			0.36

Experiment 2 in Table 1 shows what happens when an increasing-order ranking of the simplicity scores is used. A negative score here underscores the correctness of both the simplicity score as well as that of the reverse-ranking.

The third system, SALSA (Stand-alone Lexical Substitution Analyzer) is the only system out of the

[1]The normalization is done by dividing by the maximum value obtained for that particular resource

three that takes advantage of the context provided with the data set. It builds all possible 3-grams from the context, replacing the target word one-by-one by a substitute candidate (and inflections of the substitute candidates). It then sums their frequency counts in the Web1T corpus and assigns the sum to the simplicity score of a particular synonym. The synonyms can then be reverse-ranked.

5 System Standings and Discussion

For the test data, Table 2 depicts the system standings, separated by statistical significance.

Table 2: Test data system scores

Rank	Team ID	System ID	Score
1	WLV-SHEF	SimpLex	0.496
2	baseline	Sim Freq	0.471
2	**UNT**	**SimpRank**	**0.471**
2	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	No Change	0.106
9	baseline	Rand	0.013
10	UNT	SaLSA	-0.082

Surprisingly, the systems SIMPRANK and SIMPRANKLIGHT, which do not use the contexts provided, score much better than SALSA, which does use the contexts. Apparently simplicity is rather a statistical concept even for humans (the annotators for the gold standard) and not a contextual one. Also surprisingly, SIMPRANKLIGHT, which does not use Google Web1T data, performs extremely well and within 0.02 of the raw scores.

What is also surprising is the inability of all-but-one systems to beat the baseline of using simple frequency counts from Web1T, which is in turn based entirely on statistical counts and does not take the context into account.

A major contribution of this paper is the discovery that other, lighter, free resources work just as well as the expensive (in money, time and space) Web1T data when it comes to identifying which word is simple and which one is not.

6 Future Work

I plan to extend this experiment by performing ablation studies of all the individual features, playing with new features, and also performing machine learning experiments to see if supervised experiments are a better way of solving the problem of lexical simplicity ranking.

References

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *In Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Diana McCarthy, Falmer East Sussex, and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *In Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of the International Conference RANLP-2009*, pages 404–410, Borovets, Bulgaria, September. Association for Computational Linguistics.

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 (SemEval 2012)*, Montreal, Canada.

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*, pages 365–368.