UOW-SHEF: SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features

Sujay Kumar Jauhar
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street, Wolverhampton
WV1 1SB, UK
Sujay.KumarJauhar@wlv.ac.uk

Lucia Specia
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP, UK
L.Specia@dcs.shef.ac.uk

Abstract

This paper describes SimpLex,[1] a Lexical Simplification system that participated in the English Lexical Simplification shared task at SemEval-2012. It operates on the basis of a linear weighted ranking function composed of context sensitive and psycholinguistic features. The system outperforms a very strong baseline, and ranked first on the shared task.

1 Introduction

Lexical Simplification revolves around replacing words by their simplest synonym in a context aware fashion. It is similar in many respects to the task of Lexical Substitution (McCarthy and Navigli, 2007) in that it involves elements of selectional preference on the basis of a central predefined criterion (simplicity in the current case), as well as sensitivity to context.

Lexical Simplification envisages principally a human target audience, and can greatly benefit children, second language learners, people with low literacy levels or cognitive disabilities, and in general facilitate the dissemination of knowledge to wider audiences.

We experimented with a number of features that we posited might be inherently linked with textual simplicity and selected the three that seemed the most promising on an evaluation with the trial dataset. These include contextual and psycholinguistic components. When combined using an SVM ranker to build a model, such a model provides results that offer a statistically significant improvement over a very strong context-independent baseline. The system ranked first overall on the Lexical Simplification task.

[1] Developed by co-organizers of the shared task

2 Related Work

Lexical Simplification has received considerably less interest in the NLP community as compared with Syntactic Simplification. However, there are a number of notable works related to the topic.

In particular Yatskar et al. (2010) leverage the relations between Simple Wikipedia and English Wikipedia to extract simplification pairs. Biran et al. (2011) extend this base methodology to apply lexical simplification to input sentences. De Belder and Moens (2010), in contrast, provide a more general architecture for the task, with scope for possible extension to other languages.

These studies and others have envisaged a range of different target user groups including children (De Belder and Moens, 2010), people with low literacy levels (Aluisio et al., 2008) and aphasic readers (Carroll et al., 1998).

The current work differs from previous research in that it envisages a stand-alone lexical simplification system based on linguistically motivated and cognitive principles within the framework of a shared task. Its core methodology remains open to integration into a larger Text Simplification system.

3 Task Setup

The English Lexical Simplification shared task at SemEval-2012 (Specia et al., 2012) required sys-

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 477–481,
Montréal, Canada, June 7-8, 2012.

tems to rank a number of candidate substitutes (which were provided beforehand) based on their simplicity of usage in a given context. For example, given the following context with an empty placeholder, and its candidate substitutes:

> **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.
>
> **Candidates:** {clever} {smart} {intelligent} {bright}

a system is required to produce a ranking, e.g.:

> **System:** {intelligent} {bright} {clever, smart}

Note that ties were permitted and that all candidates needed to be included in the system rankings.

4 The SimpLex Lexical Simplification System

In an approach similar to what Hassan et al. (2007) used for Lexical Substitution, SimpLex ranks candidates based on a weighted linear scoring function, which has the generalized form:

$$s\left(c_{n,i}\right) = \sum_{m \in M} \frac{1}{r^m\left(c_{n,i}\right)}$$

where $c_{n,i}$ is the candidate substitute to be scored, and each r^m is a standalone ranking function that attributes to each candidate its rank based on its uniquely associated features. Based on this scoring, candidates for context are ranked in descending order of scores.

In the development of the system we experimented with a number of these features including ranking based on word length, number of syllables, scoring with a 2-step cluster and rank architecture, latent semantic analysis, and average point-wise mutual information between the candidate and neighboring words in the context.

However, the features which were intuitively the simplest proved, in the end, to give the best results. They were selected based on their superior performance on the trial dataset and their competitiveness with the strong Simple Frequency baseline. These stand-alone features are described in what follows.

4.1 Adapted N-Gram Model

The motivation behind an n-gram model for Lexical Simplification is that the task involves an inherent WSD problem. This is because the same word may be used with different senses (and consequently different levels of complexity) in different contexts.

A blind application of n-gram frequency searching on the shared task's dataset, however, gives suboptimal results because of two main factors:

1. Inconsistently lemmatized candidates.
2. Blind replacement of even correctly lemmatized forms in context producing ungrammatical results.

We infer the correct inflection of all candidates for a given context based on the appearance of the original target word (which is also one of the candidate substitutes) in context. To do this we run a part-of-speech (POS) tagger on the source text and note the POS of the target word. Then handcrafted rules are used to correctly inflect the other candidates based on this POS tag.

To resolve the issue of ungrammatical textual output, we further use a simple approach of *popping* words in close proximity to the placeholder and performing n-gram searches on all possible query combinations. Take for instance the following example:

> **Context:** He was _____ away.
>
> **Candidates:** {going} {leaving}

where "going" is evidently the original word in context, but "leaving" has also been suggested as a substitute (there are many such cases in the datasets). One of the possible outcomes of *popping* context words leads to the correct sequence for the latter substitute, i.e. "He was **leaving**" with the word "away" having been *popped*.

The rationale behind this approach is that if one of the combinations is grammatically correct, the number of n-gram hits it returns will far exceed those returned by ungrammatical ones.

The n-gram ($2 \leq n \leq 5$) searches are performed on the Google Web 1T corpus (Brants and Franz, 2006), and the number of hits is weighted by the length of the n-gram search (such that longer sequences obtain higher weight). This may seem like

a simplistic approach, especially when the candidate words appear in long-distance dependency relations to other parts of the sentence. However, it should be noted that since the Web 1T corpus only consists of n-grams with $n \leq 5$, structures that contain longer dependencies than this are in any case not considered, and hence do not interfere with local context.

4.2 Bag-of-Words Model

The limitations of performing queries on the Google Web 1T are that n-grams hits must be in strict linear order of appearance. To overcome this difficulty, we further mimic the functioning of a bag-of-words model by taking all possible ordering of words of a given n-gram sequence. This approach, to some extent, gives the possibility of observing co-occurrences of candidate and context words in various orderings of appearance. This results in a number of inadequate query strings, but possibly a few (as opposed to one in a linear n-gram search) good word orderings with high hits as well.

As with the previous model, only n-grams with $2 \leq n \leq 5$ are taken. For a given substitute the total number of hits for all possible queries involving that substitute are summed (with each hit being weighted by the length of its corresponding query in words). To obtain the final score, this sum is normalized by the actual number of queries.

4.3 Psycholinguistic Feature Model

The MRC Psycholinguistic Database (Wilson, 1988) and the Bristol Norms (Stadthagen-Gonzalez and Davis, 2006) are knowledge repositories that associate scores to words based on a number of psycholinguistic features. The ones that we felt were most pertinent to our study are:

1. Concreteness - the level of abstraction associated with the concept a word describes.

2. Imageability - the ability of a given word to arouse mental images.

3. Familiarity - the frequency of exposure to a word.

4. Age of Acquisition - the age at which a given word is appropriated by a speaker.

We combined both databases and compiled a single resource consisting of all the words from both sources that list at least one of these features. It may be noted that these attributes were compiled in similar fashion in both databases and were normalized to the same scale of scores falling in the range of 100 to 700.

In spite of a combined compilation, the coverage of the resource was poor, with more than half the candidate substitutes on both trial and test sets simply not being listed in the databases. To overcome this difficulty we introduced a fifth *frequency* feature that essentially simulates the "Simple Frequency" baseline, [2] but with scores that were normalized to the same scale of the other psycholinguistic features.

This composite of features was used in a linear weighted function with weights tuned to best performance values on the trial dataset. This function sums the weighted scores for each candidate, and normalizes this sum by the number of non-zero features (in the worst-case scenario, – when no psycholinguistic features are found – the scorer is equivalent to the "Simple Frequency" baseline). It is interesting to note that the frequency feature did not dominate the linear combination; rather there was a nice interplay of features with Concreteness, Imageability, Familiarity, Age of Acquisition and Simple Frequency being weighted (on a scale of -1 to +1) as 0.72, -0.22, 0.87, 0.36 and 0.36, respectively.

4.4 Feature Combination

We combined the three standalone models using the ranking function of the SVM-light package (Joachims, 2006) for building SVM rankers. The parameters of the SVM were tuned on the trial dataset, which consisted of only 300 example contexts. To avoid overfitting, instead of taking the single best parameters, we took parameter values that were the average of the top 10 distinct runs.

It may be noted that the resulting model makes no attempt to tie candidates, although actual ties may be produced by chance. But since ties are rarely used in the gold standard for the trial dataset, we reasoned that this should not affect the system performance in any significant way.

[2] The "Simple Frequency" baseline scores each substitute based on the number of hits it produces in the Google Web 1T

	bline-SFreq	w-ln	n-syll	psycho	a-n-gram	b-o-w	pmi	lsa	SimpLex
Trial	0.398	0.176	0.118	0.388	0.397	0.395	0.340	0.089	–
Test	0.471	0.236	0.163	0.432	0.460	0.460	0.404	0.054	0.496

Table 1: Comparison of Models' Scores

5 Results and Discussion

The results of the SimpLex system trained and tuned on the trial set, in comparison with the Simple Frequency baseline and the other stand-alone features we experimented with are presented in Table 1. The scores are computed through a version of the Kappa index over pairwise rankings, and therefore represent the average agreement between the system and the gold-standard annotation in the ranking of pairs of candidate substitutes.

Table 1 shows that while in isolation the features are unable to beat the Simple Frequency model, together they form a combination which outperforms the baseline. The improvement of SimpLex over the other models is statistically significant (statistical significance was established using a randomization test with 1000 iterations and p-value ≤ 0.05).

We believe that the reason why the context aware features were still unable to score better than the context-independent baseline is the isolated focus on simplifying a single target word. People tend to produce language that contains words of roughly equal levels of complexity. Hence in some cases the surrounding context, instead of helping to disambiguate the target word, introduces further noise to queries, especially when its individual component words have skewed complexity factors. A simultaneous simplification of all the content words in a context could be a possible solution to this problem.

As an additional experiment to assess the importance of the size of the training data in our simplification system, we pooled together the trial and test datasets, and ran several iterations of the combination algorithm with a regular increment of number of training examples and noted the effects it produced on eventual score. Three hundred examples were apportioned consistently to a test set to maintain comparability between experiments. Note that this time, no optimization of the SVM parameters was made. The results were inconclusive, and contrary to expectation, revealed that there is no general improvement with additional training data. This could be because of the difficulty of the learning problem, for which the scope of the combined dataset is still very limited. A more detailed study with a corpus that is orders of magnitude larger than the current one may be necessary to establish conclusive evidence.

6 Conclusion

This paper presented our system SimpLex which participated in the English Lexical Simplification shared-task at SemEval-2012 and ranked first out of 9 participating systems.

Our findings showed that while a context agnostic frequency approach to lexical simplification seems to effectively model the problem of assessing word complexity to a relatively decent level of accuracy, as evidenced by the strong baseline of the shared task, other elements, such as interplay of context awareness with humanly perceived psycholinguistic features can produce better results, in spite of very limited training data.

Finally, a more global approach to lexical simplification that concurrently addresses all the words in a context to normalize simplicity levels, may be a more realistic proposition for target applications, and also help context aware features perform better.

Acknowledgment

This work was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT program.

References

Sandra M. Aluisio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceeding of the eighth ACM sym-*

posium on Document engineering, DocEng '08, pages 240–248, Sao Paulo, Brazil. ACM.

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, USA, June. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. The google web 1t 5-gram corpus version 1.1 ldc2006t13.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of AAAI - 98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Madison, Wisconsin, July.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on Accessible Search Systems*, pages 19–26. ACM, July.

S. Hassan, A. Csomai, C. Banea, R. Sinha, and R. Mihalcea. 2007. Unt: Subfinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410 – 413. Association for Computational Linguistics, Prague, Czech Republic.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

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.

Lucia Specia, Sujay Kumar 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.

Hans Stadthagen-Gonzalez and Colin Davis. 2006. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38:598–605.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods*, 20:6–10.

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*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.