SB: mmSystem - Using Decompositional Semantics for Lexical Simplification

Marilisa Amoia
Department of Applied Linguistics
University of Saarland
m.amoia@mx.uni-saarland.de

Massimo Romanelli
DFKI GmBH
Saarbrcken, Germany
romanell@dfki.de

Abstract

In this paper, we describe the system we submitted to the SemEval-2012 Lexical Simplification Task. Our system (mmSystem) combines word frequency with decompositional semantics criteria based on syntactic structure in order to rank candidate substitutes of lexical forms of arbitrary syntactic complexity (one-word, multi-word, etc.) in descending order of (cognitive) simplicity. We believe that the proposed approach might help to shed light on the interplay between linguistic features and lexical complexity in general.

1 Introduction

Lexical simplification is a subtask of the more general text simplification task which attempts at reducing the cognitive complexity of a text so that it can be (better) understood by a larger audience. Text simplification has a wide range of applications which includes applications for the elderly, learners of a second language, children or people with cognitive deficiencies, etc.

Works on text simplification mostly focus on reducing the syntactic complexity of the text (Siddharthan, 2011; Siddharthan, 2006) and only little work has addressed the issue of lexical simplification (Devlin, 1999; Carroll et al., 1999).

The Lexical Simplification Task (Specia et al., 2012) proposed within the SemEval-2012 is the first attempt to explore the nature of the lexical simplification more systematically. This task requires participating systems, given a context and a target word, to automatically generate a ranking of substitutes, i.e. lexical forms conveying similar meanings to the target word, such that cognitively simpler lexical forms are ranked higher than more difficult ones.

In this paper, we describe the system we submitted to the SemEval-2012 Lexical Simplification Task. In order to rank the candidate substitutes of a lexical form in descending order of simplicity, our system (mmSystem) combines word frequency with decompositional semantics criteria based on syntactic structure. The mmSystem achieved an average ranking if compared with the other participating systems and the baselines. We believe that the approach proposed in this paper might help to shed light on the interplay between linguistic features and cognitive complexity in general.

2 The Lexical Simplification Task

The SemEval-2012 Lexical Simplification Task requires participating systems to automatically generate a ranking of lexical forms conveying similar meanings on cognitive simplicity criteria and can be defined as follows. Given a short text C called the context and which generally corresponds to a sentence, a target word T and a list L_S of candidate substitutes for T, i.e. a list of quasi-synonyms of the target word, the task for a system consists in providing a ranking on L_S such that the original list of substitutes is sorted over simplicity, from the cognitively simplest to the cognitively most difficult lexical form.

As the examples from (1) to (3) show, the Lexical Simplification Task includes substitutes of different syntactic complexity which might vary from simple one-word substitutes as in (1) (the lexical forms that

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 482–486,
Montréal, Canada, June 7-8, 2012.

can function as substitutes include content words, i.e. nouns (n), verbs (v), adjectives (a) and adverbs (r)) to collocations, negated forms as in (2) or even definition-like paraphrases as for instance *wind* and *knock the breath out of* in example (3).

(1)

C: He suggested building an experimental hypertext 'web' for the *worldwide.a* community of physicists who used CERN and its publications.
T: *worldwide.a*
$\mathbf{L_S}$: *worldwide, global, international*

(2)

C: Go to hell! she remembers Paul yelling at her *shortly.r* after their wedding.
T: *shortly.r*
$\mathbf{L_S}$: *soon, a little, just, almost immediately, shortly, not long*

(3)

C: Now however she was falling through that skylight, the strong dark figure that had appeared out of nowhere falling through with her, his arms tightly entwined about her, his shoulder having *winded.v* her.
T: *winded.v*
$\mathbf{L_S}$: *knock her breathless, knock the wind out of, choke, wind, knock the breath out of, knock the air out of*

The organizers of the Lexical Simplification Task provide a corpus of 300 trial and 1710 test sentences defining the context of the target word and the associated list of candidate substitutes. To produce a gold standard, 5 human annotators manually ranked the list of substitutes associated to each context. Finally, a scoring algorithm is provided for computing agreement between the output of the system and the manually ranked gold standard. The scoring algorithm is based on the Kappa measure for inter-annotator agreement.

3 The mmSystem

Our aim by participating in the SemEval-2012 Lexical Simplification Task (Task 1) was to investigate the nature of lexical simplicity/complexity and to identify the linguistic features that are responsible for it. The system we have developed is a first step in this direction. The idea behind our framework is the following. We build on previous work (Devlin, 1999; Carroll et al., 1999) that approximate simplicity with word frequency, such that the cognitively simpler lexical form is the one that is more frequent in the language. While this definition might easily apply to one-word substitutes or collocations, it poses some problems in the case of multi-word-expressions or of syntactically more complex lexical forms (e.g. definition like paraphrases) like those proposed in the substitute lists in the SemEval-2012 Task 1.

Our approach builds on the baseline definition of simplicity based on word frequency and integrates it with (de)compositional semantics considerations. Therefore, in order to operationalize the notion of simplicity in our system we adopt different strategies depending on the syntactic complexity of the lexical form that forms the substitute.

- In the case of one-word substitutes or common collocations we use the frequency associated by WordNet (Fellbaum, 1998) to the lexical form as a metric to rank the substitutes, i.e. the substitute with the highest frequency is ranked higher. For instance, the lexical item *intelligent* is ranked lower than *clever* as it has a lower frequency in the language (as defined in WordNet).

- In the case of multi-words or syntactic complex substitutes, we apply so-called *relevance rules*. Those are based on (de)compositional semantic criteria and attempt to identify a unique content word in the substitute that better approximates the whole lexical form. Thus, we assign to the whole lexical form the frequency associated to this most relevant content word and use it for ranking the whole substitute. For instance, relevance rules assign to multi-word substitutes such as *most able* or *not able* the same frequency, and namely that associated with the content word *able*.

3.1 Implementation

In this section we describe in more details the implementation of the mmSystem. The system design can be summarized as follows.

Step 1: POS-Tagging In the first step, context and the associated substitutes are parsed[1] so to obtain a flat representation of their syntax. Basically at this level, we collect Part-Of-Speech information for all content words in the context as well as in the substitute list.

Step 2: Relevance Rules In the second step, depending on the syntactic representation of the substitutes, the system selects a relevance rule that identifies the one-word lexical form that will be used for representing the meaning of the whole substitute.

Step 3: Word Sense Tagging The system applies word sense tagging and assigns a WordNet sense to the target words and their candidate substitutes. In this step, we rely on the SenseRelate::TargetWord package (Patwardhan et al., 2005) and use the Lesk algorithm (Lesk, 1986) for word sense disambiguation.

Step 4: Substitute Ranking Following (Carroll et al., 1999) that pointed out that rare words generally have only one sense, in order to associate a frequency index to each candidate substitute (w_i), we use the number of senses associated by WordNet to a lexical item of a given part of speech, as an approximation of its frequency (f_i). Further, we extract from WordNet the frequency of the word sense (f_{wns_i}) associated to the lexical item w_i at step 3. Words not found in WordNet it assigned a null frequency ($f_i = 0$, $f_{wns_i} = 0$). Finally, we rank the substitute in the following way:

- if $f_1 \neq f_2$
  $w_1 < w_2$, if $f_1 > f_2$ and
  $w_2 < w_1$ otherwise,
- else if $f_1 = f_2$
  $w_1 < w_2$, if $f_{wns_1} > f_{wns_2}$ and
  $w_2 < w_1$ otherwise.

Input:
Sentence 993: "It is *light.a* and easy to use." Substitutes: portable;unheavy;not heavy;light
Step 1: POS-Tagging portable#A; unheavy#A; not#Neg heavy#A; light#A
Step 2: Relevance Rules portable#A; unheavy#A; heavy#A#; light#A
Step 3: WSD portable#A#wns:2; unheavy#A#wns:?; heavy#A#wns:2; light#A#wns:25
Step 4: Ranking portable#f:2; unheavy#f:0; heavy#f:27; light#f:25 not heavy $<$ light $<$ portable $<$ unheavy
Gold Ranking: light $<$ not heavy $<$ portable $<$ unheavy

Table 1: Example of mmSystem processing steps.

Table 1 shows an example of data processing.

3.2 Relevance Rules

Relying on previous work on compositional semantics of multi-word-expression (Reddy et al., 2011; Venkatapathy and Joshi, 2005; Baldwin et al., 2003) we defined a set of hand-written rules to assign the relevant meaning to a complex substitute. Relevance rules are used to decompose the meaning of a complex structure and identify the most relevant word conveying the semantics of the whole, so that the frequency associated to the whole lexical form is approximated by the frequency of this most relevant form:

- a one-word lexical item is mapped to itself, e.g. $run.v \rightarrow run.v$
- a multi-word lexical form including only one content word is mapped to this content word, e.g. $not.Neg\ nice.a \rightarrow nice.a$ or $be.Cop\ able.a \rightarrow able.a$
- in the case of a multi-word lexical item including more than one content word, we take into account the syntactic structure of the lexical item and apply heuristics to decide which content word is more relevant for the meaning of the whole. The heuristics we used are based on the empirical analysis of the trial data set provided by the Task 1 organizers that contains

[1] We used the Stanford Parser (Klein and Manning, 2003).

about 300 contexts. As an example consider a lexical item including a verb construction with structure $V_1 + to + V_2$ that is mapped by our rules to the second verb form V_2, e.g. $try.V_1\ to\ escape.V_2 \rightarrow escape.V_2$.

Table 2 shows some examples of relevance rules defined in the mmSystem.

Syntax	Example	R_Form
V + Prep	engage for	V
Cop + Adj	be able	Adj
Cop + V	be worried	V
Adv + V	anxiously anticipate	Adv
Adj+N	adnormal growth	Adj
N1 + N2	death penalty	N1
N1 + PrepOf + N2	person of authority	N2
V+N	take notice	N
V1+to+V2	try to escape	V2

Table 2: Example of relevance rules.

These relevance rules allow for a preliminary investigation of the nature of lexical complexity. For instance, we found that in many cases, it is the modifying element of a complex expression that is responsible for a shift in lexical complexity:

(4) a. lie<say **falsely**<say **untruthfully**

b. sample< **typical** sample < **representative** sample

4 Results

The Task 1 overall result can be found in (Specia et al., 2012). The mmSystem achieved an average ranking (score=0.289) if compared with the other participating systems and the baselines that corresponds to an absolute inter-annotator agreement between system output and golden-standard around 66%. Interestingly none of the systems achieved an absolute agreement higher than 75% in this task. This confirms that lexical simplification still remains a difficult task and that the nature of the phenomena underlying it should be better explored.

Table 3 shows the performance of our system per syntactic category. The values are a bit higher than in the official results of Task 1 as the system version used for submission was buggy, however the ranking of our system with respect to the other participating systems remains the same. Interestingly, the best score were achieved for adverbs (0.352) and adjectives (0.342). This can be explained with the fact that the decompositional semantics of these category is better accounted for by our rules.

The relative low performance achieved by the mmSystem can be explained by the fact that our rules only select one content word and use its frequency for ranking. This metric alone is clearly not enough to explain all cases of lexical simplification. As an example of the complexity of this issue, consider the interplay of negation and compositional semantics: The negation of a very frequent verb form might not be so simple to understand as its antonym, e.g. *don't, not remember/forget* vs. *omit to, fail to remember/forget*. We believe, that a more systematic analysis of the lexical semantics involved in lexical simplicity might improve the performance of the system.

	Noun	**Verb**	**Adj**	**Adv**	**TOT**
cAgr:	0.5	0.5	0.5	0.5	0.5
aAgr:	0.658	0.658	0.671	0.676	0.665
Score:	0.316	0.315	0.342	0.352	0.329

Table 3: mmSystem scores per syntactic category. In the table cAgr represents the agreement by chance, aAgr is the absolute inter-annotator agreement between system output and gold ranking and score is the normalized system score. These values corresponds to P(A) and P(E) observed in the data.

5 Conclusion

In this paper we presented the mmSystem for lexical simplification we submitted to the SemEval-2012 Task 1. The system combines simplification strategies based on word frequency with decompositional semantic criteria. The mmSystem achieved an average performance. The aim of our work was in fact a preliminary investigation of the interplay between (de)compositional semantics and lexical or cognitive simplicity in general. Doubtlessly much remain to be done in order to provide a more efficient formalization of such effects. In future work, we want to perform a wider corpus analysis and study the impact of other linguistic features such as lexical semantics on lexical simplicity.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 89–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL*, pages 269–270.

S. Devlin. 1999. *Simplifying natural language for aphasic readers*. Ph.D. thesis, University of Sunderland, UK.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOV '86*.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2005. Senserelate::targetword - a generalized framework for word sense disambiguation. In *Proceedings of the Demonstration and Interactive Poster Session of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 73–76, Ann Arbor, MI.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the International Joint Conference on Natural Language Processing 2011 (IJCNLP-2011)*, Thailand.

Advaith Siddharthan. 2006. Syntactic simplification ant text cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2011. Text simplification using typed dependencies: A comparision of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on NLG*.

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.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 899–906, Stroudsburg, PA, USA. Association for Computational Linguistics.