Unsupervised Sentence Simplification Using Deep Semantics
Shashi Narayan, Claire Gardent

To cite this version:

HAL Id: hal-01623829
https://inria.hal.science/hal-01623829
Submitted on 25 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

We present a novel approach to sentence simplification which departs from previous work in two main ways. First, it requires neither hand written rules nor a training corpus of aligned standard and simplified sentences. Second, sentence splitting operates on deep semantic structure. We show (i) that the unsupervised framework we propose is competitive with four state-of-the-art supervised systems and (ii) that our semantic based approach allows for a principled and effective handling of sentence splitting.

1 Introduction

Sentence simplification maps a sentence to a simpler, more readable one approximating its content. As has been argued in (Shardlow, 2014), sentence simplification has many potential applications. It is useful as a preprocessing step for a variety of NLP systems such as parsers and machine translation systems (Chandrasekar et al., 1996), summarisation (Knight and Marcu, 2000), sentence fusion (Filippova and Strube, 2008) and semantic role labelling (Vickrey and Koller, 2008). It also has wide ranging potential societal applications as a reading aid for people with aphasia (Carroll et al., 1999), for low literacy readers (Watanabe et al., 2009) and for non native speakers (Siddharthan, 2002).

In this paper, we present a novel approach to sentence simplification which departs from previous work in two main ways. First, it requires neither hand written rules nor a training corpus of aligned standard and simplified sentences. Instead, we exploit non aligned Simple and English Wikipedia to learn the probability of lexical simplifications, of the semantics of simple sentences and of optional phrases i.e., phrase which may be deleted when simplifying. Second, sentence splitting is semantic based. We show (i) that our unsupervised framework is competitive with four state-of-the-art systems and (ii) that our semantic based approach allows for a principled and effective handling of sentence splitting.

2 Related Work

Earlier work on sentence simplification relied on handcrafted rules to capture syntactic simplification e.g., to split coordinated and subordinated sentences into several, simpler clauses or to model e.g., active/passive transformations (Siddharthan, 2002; Chandrasekar and Srinivas, 1997; Canning, 2002; Siddharthan, 2011; Siddharthan, 2010). While these hand-crafted approaches can encode precise and linguistically well-informed syntactic transformations, they do not account for lexical simplifications and their interaction with the sentential context. Siddharthan and Mandya (2014) therefore propose an approach where hand-crafted syntactic simplification rules are combined with lexical simplification rules extracted from aligned English and simple English sentences, and revision histories of Simple Wikipedia.

Using the parallel dataset formed by Simple English Wikipedia (SWKP) and traditional English Wikipedia (EWKP), further work has focused on developing machine learning approaches to sentence simplification.

Zhu et al. (2010) constructed a parallel Wikipedia corpus (PWKP) of 108,016/114,924 complex/simple sentences by aligning sentences from EWKP and SWKP and used the resulting bitext to train a simplification model inspired by syntax-based machine translation (Yamada and Knight, 2001). Their simplification model encodes the probabilities for four rewriting operations on the parse tree of an input sen-

\[ http://simple.wikipedia.org \]
\[ http://en.wikipedia.org \]
tences namely, substitution, reordering, splitting and deletion. It is combined with a language model to improve grammaticality and the decoder translates sentences into simpler ones by greedily selecting the output sentence with highest probability.

Using both the PWKP corpus developed by Zhu et al. (2010) and the edit history of simple Wikipedia, Woodsend and Lapata (2011) learn a quasi synchronous grammar (Smith and Eisner, 2006) describing a loose alignment between parse trees of complex and of simple sentences. Following Dras (1999), they then generate all possible rewrites for a source tree and use integer linear programming to select the most appropriate simplification. They evaluate their model on the same dataset used by Zhu et al. (2010) namely, an aligned corpus of 100/131 EWKP/SWKP sentences.

Wubben et al. (2012), Coster and Kauchak (2011) and Xu et al. (2016) saw simplification as a monolingual translation task where the complex sentence is the source and the simpler one is the target. To account for deletions, reordering and substitution, Coster and Kauchak (2011) trained a phrase based machine translation system on the PWKP corpus while modifying the word alignment output by GIZA++ in Moses to allow for null phrasal alignments. In this way, they allow for phrases to be deleted during translation. Similarly, Wubben et al. (2012) used Moses and the PWKP data to train a phrase based machine translation system augmented with a post-hoc reranking procedure designed to rank the output based on their dissimilarity from the source sentence. Unlike Wubben et al. (2012) and Coster and Kauchak (2011) who used machine translation as a black box, Xu et al. (2016) proposed to modify the optimization function of SMT systems by tuning them for the sentence simplification task. However, in their work they primarily focus on lexical simplification.

Finally, Narayan and Gardent (2014) present a hybrid approach combining a probabilistic model for sentence splitting and deletion with a statistical machine translation system trained on PWKP for substitution and reordering.

Our proposal differs from all these approaches in that it does not use the parallel PWKP corpus for training. Nor do we use hand-written rules. Another difference is that we use a deep semantic representation as input for simplification. While a similar approach was proposed in (Narayan and Gardent, 2014), the probabilistic models differ in that we determine splitting points based on the maximum likelihood of sequences of thematic role sets present in SWKP whereas Narayan and Gardent (2014) derive the probability of a split from the aligned EWKP/SWKP corpus using expectation maximization. As we shall see in Section 4, because their data is more sparse, Narayan and Gardent (2014) predicts less and lower quality simplifications by sentence splitting.

3 Simplification Framework

Our simplification framework pipelines three dedicated modules inspired from previous work on lexical simplification, syntactic simplification and sentence compression. All three modules are unsupervised.

3.1 Example Simplification

Before describing the three main modules of our simplification framework, we illustrate its working with an example. Figure 1 shows the input semantic representation associated with sentence (1C) and illustrates the successive simplification steps yielding the intermediate and final simplified sentences shown in (1S1-S).

(1) C. In 1964 Peter Higgs published his second paper in Physical Review Letters describing Higgs mechanism which predicted a new massive spin-zero boson for the first time.

S1 (Lex Simp). In 1964 Peter Higgs wrote his second paper in Physical Review Letters explaining Higgs mechanism which predicted a new massive elementary particle for the first time.

S2 (Split). In 1964 Peter Higgs wrote his second paper in Physical Review Letters explaining Higgs mechanism. Higgs mechanism predicted a new massive elementary particle for the first time.

S (Deletion). In 1964 Peter Higgs wrote his paper explaining Higgs mechanism. Higgs mechanism predicted a new elementary particle.

First, the input (1C) is rewritten as (1S1) by replacing standard words with simpler ones using the context aware lexical simplification method proposed in (Biran et al., 2011).

Splitting is then applied to the semantic representation of (1S1). Following Narayan and Gardent (2014), we use Boxer 3 (Curran et al., 2007) to map the output sentence from the lexical simplification step (here S1) to a Discourse Representation Structure (DRS, (Kamp, 1981)). The DRS for

3http://svn.ask.it.usyd.edu.au/trac/candc, Version 1.00
In 1964 Peter Higgs published his second paper in Physical Review Letters describing Higgs mechanism which predicted a new massive spin-zero boson for the first time.

In 1964 Peter Higgs wrote his second paper in Physical Review Letters explaining Higgs mechanism which predicted a new massive elementary particle for the first time.

Figure 1: Simplification of “In 1964 Peter Higgs published his second paper in Physical Review Letters describing Higgs mechanism which predicted a new massive spin-zero boson for the first time.”
$S_1$ is shown at the top of Figure 1 and a graph representation of the dependencies between its variables is shown immediately below. In this graph, each DRS variable labels a node in the graph and each edge is labelled with the relation holding between the variables labelling its end vertices. The two tables to the right of the picture show the predicates (top table) associated with each variable and the relation label (bottom table) associated with each edge. Boxer also outputs the associated positions in the complex sentence for each predicate (not shown in the DRS but shown in the graph tables). Orphan words i.e., words which have no corresponding material in the DRS (e.g., which at position 16), are added to the graph (node $O_1$) thus ensuring that the position set associated with the graph exactly generates the input sentence.

Using probabilities over sequences of thematic role sets acquired from the DRS representations of SWKP, the split module determines where and how to split the input DRS. In this case, one split is applied between $X_{11}$ (explain) and $X_{10}$ (predict). The simpler sentences resulting from the split are then derived from the DRS using the word order information associated with the predicates, duplicating or pronominalising any shared element (e.g., Higgs mechanism in Figure 1) and deleting any Orphan words (e.g., which) which occurs at the split boundary. Splitting thus derives $S_2$ from $S_1$.

Finally, deletion or sentence compression applies transforming $S_2$ into $S_3$.

### 3.2 Context-Aware Lexical Simplification

We extract context-aware lexical simplification rules from EWKP and SWKP using the approach described by Biran et al. (2011). The underlying intuition behind these rules is that the word $C$ from EWKP can be replaced with a word $S$ from SWKP if $C$ and $S$ share similar contexts (ten token window) in EWKP and SWKP respectively. Given an input sentence and the set of simplification rules extracted from EWKP and SWKP, we then consider all possible $(C, S)$ substitutions licensed by the extracted rules and we identify the best combination of lexical simplifications using dynamic programming and rule scores which capture the adequacy, in context, of each possible substitution.

### 3.3 Sentence Splitting

A distinguishing feature of our approach is that splitting is based on deep semantic representations rather than phrase structure trees – as in (Zhu et al., 2010; Woodsend and Lapata, 2011) – or dependency trees – as in (Siddharthan and Mandya, 2014).

While Woodsend and Lapata (2011) report learning 438 splitting rules for their simplification approach operating on phrase structure trees Siddharthan and Mandya (2014) defines 26 hand-crafted rules for simplifying apposition and/or relative clauses in dependency structures and 85 rules to handle subordination and coordination.

In contrast, we do not need to specify or to learn complex rewrite rules for splitting a complex sentence into several simpler sentences. Instead, we simply learn the probability of sequences of thematic role sets likely to cooccur in a simplified sentence.

The intuition underlying our approach is that:

Semantic representations give a clear handle on events, on their associated roles sets and on shared elements thereby facilitating both the identification of possible splitting points and the reconstruction of shared elements in the sentences resulting from a split.

For instance, the DRS in Figure 1 makes clear that sentence $(1S_1)$ contains 3 main events and that Higgs mechanism is shared between two propositions.

To determine whether and where to split the input sentence, we use a probabilistic model trained on the DRSs of the Simple Wikipedia sentences and a language model also trained on Simple Wikipedia. Given the event variables contained in the DRS of the input sentence, we consider all possible splits between subsequences of events and choose the split(s) with maximum split score. For instance, in the sentence shown in Figure 1, there are three event variables $X_3$, $X_{10}$ and $X_{11}$

---

1. The DRS to graph conversion goes through several preprocessing steps: the relation $mn$ is inverted making modifier noun (higgs) dependent of modified noun (mechanism), named and timex are converted to unary predicates, e.g., named($x$, peter) is mapped to peter($x$) and timex($x$) = 1964 is mapped to 1964($x$); and nodes are introduced for orphan words (e.g., which).


3. For more details on the lexical simplification process, we refer the reader to Biran et al. (2011).
in the DRS. So we will consider 5 split possibilities namely, no split \(\{X_3, X_{10}, X_{11}\}\), two splits resulting in three sentences describing an event each \(\{X_3\}, \{X_{10}\}, \{X_{11}\}\) and one split resulting in two sentences describing one and two events respectively (i.e., \(\{X_3, X_{10}, X_{11}\}\), \(\{X_3, X_{10}\}, \{X_{11}\}\) and \(\{X_{10}, X_3, X_{11}\}\)). The split \(\{X_{10}, X_3, X_{11}\}\) gets the maximum split score and is chosen to split the sentence \(\{S_1\}\) producing the sentences \(\{S_2\}\).

<table>
<thead>
<tr>
<th>Semantic Pattern</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>({\text{agent, patient}})</td>
<td>0.069</td>
</tr>
<tr>
<td>({\text{agent, in, in, patient}})</td>
<td>0.002</td>
</tr>
<tr>
<td>({\text{agent, patient}, \text{agent, in, in, patient}})</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Table 1: Split Feature Table (SFT) showing some of the semantic patterns from Figure 1.

Formally, the split score \(P_{\text{split}}\) associated with the splitting of a sentence \(S\) into a sequence of sentences \(s_1 \ldots s_n\) is defined as:

\[
P_{\text{split}} = \frac{1}{n} \sum_{s_i} \frac{L_{\text{split}}}{L_{\text{split}} + |L_{\text{split}} - L_s|} \times \text{lm}_{s_i} \times \text{SFT}_{s_i}
\]

where \(n\) is the number of sentences produced after splitting; \(L_{\text{split}}\) is the average length of the split sentences \(L_{\text{split}} = \frac{L_S}{n}\) where \(L_S\) is the length of the sentence \(S\); \(L_{s_i}\) is the length of the sentence \(s_i\); \(\text{lm}_{s_i}\) is the probability of \(s_i\) given by the language model and \(\text{SFT}_{s_i}\) is the likelihood of the semantic pattern associated with \(s_i\). The Split Feature Table (SFT, Table 1) is derived from the corpus of DRSs associated with the SWKP sentences and the counts of sequences of thematic role sets licenced by the DRSs of SWKP sentences. Intuitively, \(P_{\text{split}}\) favors splits involving frequent semantic patterns (frequent sequences of thematic role sets) and sub-sentences of roughly equal length. This way of semantic pattern based splitting also avoids over-splitting of a complex sentence.

### 3.4 Phrasal Deletion

Following Filippova and Strube (2008), we formulate phrase deletion as an optimization problem which is solved using integer linear programming\(^7\). Given the DRS \(K\) associated with a sentence to be simplified, for each relation \(r \in K\), the deletion module determines whether \(r\) and its associated DRS subgraphs should be deleted by maximising the following objective function:

\[
\sum_r x_{h,w} \times P(r|h) \times P(w) \quad r \notin \{\text{agent, patient, theme, eq}\}
\]

where for each relation \(r \in K\), \(x_{h,w} = 1\) if \(r\) is preserved and \(x_{h,w} = 0\) otherwise; \(P(r|h)\) is the conditional probability (estimated on the DRS corpus derived from SWKP) of \(r\) given the head label \(h\); and \(P(w)\) is the relative frequency of \(w\) in SWKP\(^8\).

Intuitively, this objective function will favor obligatory dependencies over optional ones and simple words (i.e., words that are frequent in SWKP). In addition, the objective function is subjected to constraints which ensure (i) that some deletion takes place and (ii) that the resulting DRS is a well-formed graph.

### 4 Evaluation

We evaluate our approach both globally and by module focusing in particular on the splitting component of our simplification approach.

#### 4.1 Global evaluation

The testset provided by Zhu et al. (2010) was used by four supervised systems for automatic evaluation using metrics such as BLEU, sentence length and number of edits. In addition, most recent simplification approaches carry out a human evaluation on a small set of randomly selected complex/simple sentence pairs. Thus Wubben et al. (2012), Narayan and Gardent (2014) and Siddharthan and Mandya (2014) carry out a human evaluation on 20, 20 and 25 sentences respectively.

Accordingly, we perform an automatic comparative evaluation using (Zhu et al., 2010)’s testset namely, an aligned corpus of 100/131 EWKP/SWKP sentences; and we carry out a human-based evaluation.

**Automatic Evaluation** Following Wubben et al. (2012), Zhu et al. (2010) and Woodsend and Lapata (2011), we use metrics that are directly related to the simplification task namely, the number of splits in the overall data, the number of output sentences with no edits (i.e., sentences which have not

\(^7\)In our implementation, we use lpsolve, http://sourceforge.net/projects/lpsolve.

\(^8\)To account for modifiers which are represented as predicates on nodes rather than relations, we preprocess the DRSs and transform each of these predicates into a single node subtree of the node it modifies. For example in Figure 1, the node \(X_2\) labeled with the modifier predicate second is updated to a new node \(X_2'\) dominating a child labeled with that predicate and related to \(X_2'\) by a modifier relation.
Table 2: Automatic evaluation results. Zhu, Woodsend, Wubben, Narayan are the best output of the models of Zhu et al. (2010), Woodsend and Lapata (2011), Wubben et al. (2012) and Narayan and Gardent (2014) respectively. UNSUP is our model.

Table 3: Automated Metrics for Simplification: Modular evaluation. LexSimpl-Split-Deletion is our final system UNSUP.

Table 4: Average Human Ratings for simplicity, fluency and adequacy.

---

been simplified) and the average Levenshtein distance (LD) between the system output and both the complex and the simple reference sentences. We use BLEU\(^9\) as a means to evaluate how close the systems output are to the reference corpus.

Table 2 shows the results of the automatic evaluation. The most noticeable result is that our unsupervised system yields results that are similar to those of the supervised approaches.

The results also show that, in contrast to Woodsend system which often leaves the input unsimplified (24% of the input), our system almost always modifies the input sentence (only 3% of the input are not simplified); and that the number of simplifications including a split is relatively high (49% of the cases) suggesting a good ability to split complex sentences into simpler ones.

**Human Evaluation** Human judges were asked to rate input/output pairs w.r.t. to adequacy (How much does the simplified sentence(s) preserve the meaning of the input?), to simplification (How much does the generated sentence(s) simplify the complex input?) and to fluency (how grammatical and fluent are the sentences?).

We randomly selected 18 complex sentences from Zhu’s test corpus and included in the evaluation corpus: the corresponding simple (Gold) sentence from Zhu’s test corpus, the output of our system (UNSUP) and the output of the other four systems (Zhu, Woodsend, Narayan and Wubben) which were provided to us by the system authors\(^10\). We collected ratings from 18 participants. All were either native speakers or proficient in English, having taken part in a Master taught in English or lived in an English speaking country for an extended period of time. The evaluation was done online using the LG-Eval toolkit (Kow and Belz, 2012)\(^11\) and a Latin Square Experimental Design (LSED) was used to ensure a fair distribution of the systems and the data across raters.

Table 4 shows the average ratings of the human evaluation on a scale from 0 to 5. Pairwise comparisons between all models and their statistical significance were carried out using a one-way ANOVA with post-hoc Tukey HSD tests.

---


\(^{10}\)We upload the outputs from all the systems as supplementary material with this paper.

\(^{11}\)http://www.nltg.brighton.ac.uk/research/lg-eval/
If we group together systems for which there is no significant difference (significance level: p < 0.05), our system is in the first group together with Narayan and Zhu for simplicity; in the first group for fluency; and in the second group for adequacy (together with Woodsend and Zhu). A manual examination of the results indicates that UNSUP achieves good simplicity rates through both deletion and sentence splitting. Indeed, the average word length of simplified sentences is smaller for UNSUP (26.22) than for Wubben (28.25) and Woodsend (28.10); comparable with Narayan (26.19) and higher only than Zhu (24.21).

### 4.2 Modular Evaluation

To assess the relative impact of each module (lexical simplification, deletion and sentence splitting), we also conduct an automated evaluation on each module separately. The results are shown in Table 3.

One first observation is that each module has an impact on simplification. Thus the average Levenshtein Edit distance (LD) to the source clause (complex) is never null for any module while the number of “No edit” indicates that lexical simplification modifies the input sentence in 78%, sentence splitting 49% and deletion 96% of the cases.

In terms of output quality and in particular, similarity with respect to the target clause, deletion is the most effective (smallest LD, best BLEU score w.r.t target). Further, the results for average token length indicate that lexical simplification is effective in producing shorter words (smaller average length for this module compared to the other two modules).

Predictably, combining modules yields systems that have stronger impact on the source clause (higher LD to complex, lower number of No Edits) with the full system (i.e., the system combining the 3 modules) showing the largest LD to the sources (LD to complex) and the smallest number of source sentences without simplification (3 No Edits).

### 4.3 Sentence Splitting Using Deep Semantics

To compare our sentence splitting approach with existing systems, we collected in a second human evaluation, all the outputs for which at least one system applied sentence splitting. The raters were then asked to compare pairs of split sentences produced by two distinct systems and to evaluate the quality (0:very bad to 5:very good) of these split sentences taking into account boundary choice, sentence completion and sentence reordering.

Table 5 shows the results of this second evaluation. For each system pair comparing UNSUP (A) with another system (B), the Table gives the scores and the number of splits of both systems: for the inputs on which both systems split (BOTH-AB), on which only UNSUP splits (ONLY-A) and on which only the compared system split (ONLY-B).

UNSUP achieves a better average score (ALL-A = 2.37) than all other systems (ALL-B column) except Wubben (2.73). However Wubben only achieves one split and on that sentence, UNSUP score is 4.75 while Wubben has a score of 2.73 and produces an incorrect split (cf. $S_3$ in Figure 6). UNSUP

Table 5: Pairwise split evaluation: Each row shows the pairwise comparison of the quality of splits in UNSUP and some other system. Last six columns show the average scores and number of associated split sentences. The second column (ALL-A) and the third column (ALL-B) present the quality of all splits by systems A and B respectively. The fourth column (ONLY-A) represents sentences where A splits but not B. The fifth and sixth columns represents sentences where both A and B split. The seventh column (ONLY-B) represents sentences where B splits but not A.

<table>
<thead>
<tr>
<th>System pairs</th>
<th>Average Score (number of split sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD</td>
<td>3.85(28)</td>
</tr>
<tr>
<td>Zhu</td>
<td>2.25(80)</td>
</tr>
<tr>
<td>Woodsend</td>
<td>2.08(63)</td>
</tr>
<tr>
<td>Wubben</td>
<td>2.73(1)</td>
</tr>
<tr>
<td>Narayan</td>
<td>2.09(10)</td>
</tr>
</tbody>
</table>

117
Table 6: Example Outputs for Sentence splitting with their average human annotation scores.

In sum, the unsupervised, semantic-based splitting strategy allows for a high number (49%) of good quality (2.37 score) sentence splits. Because there are less possible patterns of thematic role sets in simple sentences than possible configurations of parse/dependency trees for complex sentences, it is less prone to data sparsity than the syntax based approach. Because the probabilities learned are not tied to specific syntactic structures but to more abstract semantic patterns, it is also perhaps less sensitive to parse errors.

### 4.4 Examples from the Test Set

Table 6 shows some examples from the evaluation dataset which were selected to illustrate the workings of our approach and to help interpret the results in Table 2, 4 and 5.

S1 and S2 and S3 show examples of context-aware unsupervised lexical substitutions which are nicely performed by our system. In S1, *The array distributes data* is correctly simplified to *The array moves data* whereas Zhu’s system incorrectly simplifies this clause to *The array sells data*. Similarly, in S2, our system correctly simplifies *Papers on simulation of artificial selection* to *Papers on models of selection* while the other systems either do not simplify or simplify to *Papers on feeling*.

For splitting, the examples show two types of splitting performed by our approach namely, splitting of coordinated sentences (S1) and splitting between a main and a relative clause (S2,S3). S2 illustrates how the Woodsend system over-splits, an issue already noticed in (Siddharthan and Mandya, 2014); and how Zhu’s system predicts an incorrect split between a verb (seen) and its agent argument (by the user). Barring a parse error, such incorrect splits will not be predicted by our approach since, in our cases, splits only occur between (verbalisations of) events. S1, S2 and S3 also illustrates how our semantic based approach allows for an adequate reconstruction of shared elements.

### 5 Conclusion

A major limitation for supervised simplification systems is the limited amount of available parallel standard/simplified data. In this paper, we have shown that it is possible to take an unsupervised approach to sentence simplification which requires a large corpus of standard and simplified language but no alignment between the two. This allowed for the implementation of contextually aware substitution module; and for a simple, linguistically principled account of sentence splitting and shared element reconstruction.
6 Acknowledgements

We are grateful to Zhemin Zhu, Kristian Woodsend and Sander Wubben for sharing their data. We would like to thank our annotators for participating in our human evaluation experiments and to anonymous reviewers for their insightful comments. This research was supported by an EPSRC grant (EP/L02411X/1) and an EU H2020 grant (688139/H2020-ICT-2015: SUMMA). We also thank the French National Research Agency for funding the research presented in this paper in the context of the WebNLG project.

References


