The impact of phrases on Italian lexical simplification

Sara Tonelli, Alessio Palmero Aprosio
Fondazione Bruno Kessler
Trento, Italy
{satonelli,aprosio}@fbk.eu

Marco Mazzon
Dept. of Psychology and Cognitive Science
University of Trento
marco.mazzon@studenti.unitn.it

Abstract

English. Automated lexical simplification has been performed so far focusing only on the replacement of single tokens with single tokens, and this choice has affected both the development of systems and the creation of benchmarks. In this paper, we argue that lexical simplification in real settings should deal both with single and multi-token terms, and present a benchmark created for the task. Besides, we describe how a freely available system can be tuned to cover also the simplification of phrases, and perform an evaluation comparing different experimental settings.

Italiano. *La semplificazione lessicale automatica è stata affrontata fino ad ora dalla comunità di ricerca TAL concentrandosi sulla sostituzione di parole singole con altre parole singole. Questa modalità ha condizionato sia lo sviluppo di sistemi di semplificazione che la creazione di benchmark per la valutazione. In questo articolo, sosteniamo che la semplificazione lessicale in contesti reali debba includere sia parole singole che espressioni composte da più parole, e presentiamo un benchmark creato a questo fine. Inoltre, descriviamo come adattare un sistema disponibile per la semplificazione lessicale in modo che supporti anche la semplificazione di sintagmi, e presentiamo una valutazione confrontando diversi setting sperimentali.*

1 Introduction

Lexical simplification is a well-studied topic within the NLP community, dealing with the automatic replacement of complex terms with simpler ones in a sentence, in order to improve its clarity and readability. Thanks to the development of benchmarks (Paetzold and Specia, 2016a) and freely available tools for lexical simplification (Paetzold and Specia, 2015), a number of works have focused on this challenge, see for example the systems participating in the simplification shared task at SemEval-2012 (Specia et al., 2012). However, the task has been designed as an exercise to replace complex *single tokens* with simpler *single tokens*, and most widely used benchmarks and systems all follow this paradigm. We believe, however, that this setting covers only a limited number of lexical simplifications as they would be performed in a real scenario. In particular, we advocate the need to shift the lexical simplification paradigm from single tokens to phrases, and to develop datasets and tools that deal also with these cases. This is mainly the contribution of this work, which covers four main points:

- We analyse existing corpora of simplified texts, not specifically developed for a shared task or for system evaluation, and we measure the impact of phrases in lexical simplifications
- We modify a state-of-the-art tool for lexical simplification in order to support phrases
- We compare different strategies for phrase extraction and evaluate them over a benchmark
- We perform all the above on Italian, for which there was no lexical simplification system available.

Besides, we make freely available the first benchmark for the evaluation of Italian lexical simplification, with the goal to support research on this task and to foster the development of Italian simplification systems.

2 Corpus analysis and Benchmark creation

We first analyse existing simplification corpora in Italian to study the impact of phrases on lexical simplification. There are only two such manually created corpora, which contain different types of data but have been annotated following the same scheme: the Simpitiki corpus (Tonelli et al., 2016) and the one developed by the ItaNLP Lab in Pisa (Brunato et al., 2015). The former contains 1,163 sentence pairs[1], where one is the original sentence and the other is the simplified one. The pairs were created starting from Wikipedia edits and from documents in the public administration domain. The ItaNLP corpus, instead, contains 1,393 pairs extracted from children's stories and from educational material. Both corpora were annotated following the scheme proposed in (Brunato et al., 2015), in which simplifications were classified as *Split*, *Merge*, *Reordering*, *Insert*, *Delete* and *Transformation* (plus a set of subclasses for the *Insert*, *Delete* and *Transformation* cases). Since our goal was to isolate a benchmark of pairs containing only the lexical cases, we discarded the classes not compatible with lexical simplifications (e.g. *Delete*, *Reordering*) and then manually checked the others to identify the cases of interest. When, as in the majority of cases, a lexical simplification was present together with other simplification types, we re-wrote the target sentence in order to retain only lexical cases. For example, in the examples below, *a)* is the original sentence and *b)* is the simplified one in the Simpitiki corpus, which contains a lexical simplification of 'include' and a shift of position of 'per convenzione'. We created version *c)*, so that only the lexical simplification is present:

a) *Eurasia è il termine con cui per convenzione si definisce la zona geografica che include l'Europa e l'Asia.*
b) *Eurasia è, per convenzione, il termine con cui si definisce la zona geografica che comprende l'Europa e l'Asia.*
c) *Eurasia è il termine con cui per convenzione si definisce la zona geografica che comprende l'Europa e l'Asia.*

[1]The number is slightly different from what was reported in the original paper because the corpus was revised after the first release.

This revision process led to the creation of a benchmark with pairs extracted from the two original corpora, where only cases of lexical simplification are present[2]. Some statistics related to the benchmark are reported in Table 1. We identify four possible lexical simplification types: a single token is replaced by a single token (ST→ST), a single token is simplified through a phrase (ST→P), a phrase is simplified through a single token (P→ST), and a phrase is replaced by another phrase (P→P).

	ST→ST	ST→P	P→ST	P→P	Total
ItaNLP	369	112	139	87	707
Simpitiki	112	24	30	28	194
Total	481	136	169	115	901

Table 1: Statistics on lexical simplification benchmark (ST = Single token, P = Phrase)

We observe that the most frequent lexical simplification type is ST→ST, on which most systems and shared tasks are based. However, this simplification type covers only half of the cases included in our benchmark. This confirms the need to include cases of phrase-based simplification in the creation of benchmarks. It corroborates also the importance of developing systems for lexical simplification that support phrase replacement, so as to make them work in real settings and not only on ad-hoc test sets. Another interesting remark is that single tokens are not necessarily simpler than phrases, or vice versa: in our data, there are 136 ST→P and 169 P→ST, showing that no general rule can be applied to favour (or demote) Ps over STs.

We use the final benchmark[3], containing 901 sentence pairs, to evaluate a system for lexical simplification taking into account phrases, as described in the following Section.

3 Automated lexical simplification

In this Section we describe the experiments we carried out to perform automated lexical simplification using the benchmark presented in Section 2. We describe the tool used and how it was mod-

[2]In Simpitiki we focused only on the pairs in the public administration domain due to project constraints. We plan to include the pairs from Wikipedia in the next benchmark version.

[3]Available at `https://drive.google.com/file/d/0B4QAWZllD-egYS0yNWZ5dTdYQVE/view?usp=sharing`

ified to deal with phrases. We also detail the resources (language model and word embeddings) created for the task.

3.1 The Lexenstein system

We use Lexenstein (Paetzold and Specia, 2015), an open source tool for lexical simplification, to collect a list of candidates that should replace a given word in the text. In particular, the Paetzold generator (Paetzold and Specia, 2016b) is based on an unsupervised approach to produce simplification candidates using a context-aware word embeddings model: features used for the selection include word2vec vectors (Mikolov et al., 2013), language model created by SRILM (Stolcke, 2002), and conditional probability of a candidate given the PoS tag of the target word. So far, no evaluation on Lexestein for Italian is available.

For each complex word, five candidate replacements are first retrieved, ranked according to several features, such as n-gram frequencies and word vector similarity with the target word, and then reranked according to their average rankings (Glavaš and Štajner, 2015).

Since we wanted to test different strategies to create the embeddings (i.e. with and without phrases), we created the word/phrase vectors and the language model starting from freely available corpora (1.3 billion words in total): the Italian Wikipedia,[4] OpenSubtitles2016 (Lison and Tiedemann, 2016),[5] PAISÀ,[6] and the Gazzetta Ufficiale,[7] a collection of Italian laws. Due to the size of the data, both the corpus and the model are available upon request to the authors.

3.2 Experimental Setup

We conduct several experiments to evaluate the quality of lexical simplification when taking into account phrases (or not), and compare different strategies for phrase recognition. We compare different variants to create the embeddings and the language model (LM) that were then used by Lexenstein.

The first *baseline model* relies on the standard Lexenstein setting: word embeddings are created using the word2vec package, and the LM considers each token separately.

The first system variant (*word2phrase*) includes phrase recognition, i.e. before extracting the embeddings and creating the LM, the documents are analysed by the word2phrase module in the word2vec package. This is an implementation of the algorithm presented in (Mikolov et al., 2013), which basically identifies words that appear frequently together, and infrequently in other contexts, and treats them as single tokens (connected by an underscore).

The second system variant (*word2phrase+LemmaPos*) adds another information layer, in that each document is first lemmatized and PoS tagged using the Tint NLP Suite (Aprosio and Moretti, 2016), that works at token level; then word2phrase is run, and then the embeddings and the LM are created. In this way, we obtain so-called 'context-aware' embeddings, which is the recommended setting in (Paetzold and Specia, 2016b).

4 Evaluation

The evaluation of automated simplification is an open issue since, similar to machine translation, there may be different acceptable simplifications for a term, while a benchmark usually presents only one solution. Therefore, we perform two evaluations: the first is based on an automated comparison between Lexenstein output and the gold simplifications in the benchmark. The second is a manual evaluation aimed at scoring fluency, adequacy and simplicity of the output.

For the first evaluation, we compute the Mean Reciprocal Rank (MRR), which is usually adopted to evaluate a list of possible responses ordered by probability of correctness against a gold answer. We use this metrics because Lexenstein returns 5 possible simplifications, ranked by relevance, and with MRR it is possible to weight the response matching with the gold simplification according to its rank. In particular, MRR is computed as:

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where Q is the number of simplifications to be performed (901) and $rank_i$ is the position of the correct simplification in the rank returned by Lexenstein.

We run the system in the three configurations described in Section 3.2 on each source sentence

[4] https://it.wikipedia.org/wiki/Pagina_principale

[5] http://www.opensubtitles.org/

[6] http://www.corpusitaliano.it/

[7] http://www.gazzettaufficiale.it/

in the benchmark. The single or multi-token term to be simplified is given. If it is found in the LM, the system suggests 5 ranked simplification candidates. Otherwise, no output is given.

Results show that the baseline model, i.e. the standard Lexenstein configuration replacing only single tokens with single tokens, yields $\mathrm{MRR} = 0.036$. The one using word2phrase achieves $\mathrm{MRR} = 0.042$, while the version including also lemma and PoS information yields $\mathrm{MRR} = 0.050$. A detailed evaluation is reported in Table 2: for each of the three experimental settings, we report the number of cases in which the gold simplification matches the first ranked replacement returned by Lexenstein (*1st*), the second, the third, and so on. In the last column, we report how many times (out of 901) the rank returned by Lexenstein does not contain the gold simplification present in the benchmark.

	1st	2nd	3rd	4th	5th	none
Baseline	23	12	7	3	2	854
word2phrase	30	8	8	4	1	850
+LemmaPos	32	16	11	4	4	834

Table 2: Rank of correct simplifications returned by Lexenstein

This evaluation shows that, although limited, using word2phrase in combination with lemma and PoS information yields an improvement over the baseline. However, the informativeness of this automated simplification is limited because the cases labeled as 'none' include both wrong simplifications and correct simplifications that are not present in the benchmark. Besides, they include also cases in which the word to be simplified was not found in the LM.

In order to better understand where the approach fails, we also perform a manual evaluation. Following the standard scheme for human evaluation of automatic text simplification (Saggion and Hirst, 2017), we judge Fluency (grammaticality), Adequacy (meaning preservation) and Simplicity of lexical simplifications using a five-point Likert scale (the higher the score, the better the output). For the setting using lemma and PoS, we do not judge Fluency, since the output is lemmatized and not converted in the original form of the source term (we plan to add this in the near future). Evaluation is performed using a set of 150 sentence pairs randomly extracted from the benchmark. We introduce also this kind of evaluation in order to have a fine-grained analysis of system output. For example, in the original sentence d) (see below), 'tempestivamente' was simplified with 'periodicamente', which is grammatically correct (high Fluency) but does not preserve the meaning of the original sentence (low Adequacy).

d) Il richiedente dovrà comunicare tempestivamente l'esattezza dei recapiti forniti.

When using word2phrase without lemmatization, the average Fluency is 3.72, Adequacy is 2.60 and Simplicity is 2.95. This shows that, while PoS and form of a simplified term are generally correct also without any processing, the preservation of the meaning is a critical issue. Simplicity achieves better scores than Adequacy, but it still needs improvements. Results obtained using lemma and PoS in combination with word2phrase are slightly better, with 2.64 Adequacy and 3.01 Simplicity. In general, the above evaluations show that using word2phrase with lemma and PoS information is a promising approach to improve the performance of lexical simplification in real settings. The performance of Lexenstein could be further improved by adding other corpora to the LM and post-process the output of the system, so as to discard inconsistent simplifications, for example when a verb is simplified through an adverb. However, some linguistic phenomena like non-local dependencies cannot be addressed using this approach, and a separate strategy to simplify them should be taken into account.

5 Conclusions

In this work, we presented a first analysis of the role of phrases in Italian lexical simplification. We also introduced the adaptation of Lexenstein, an existing lexical simplification system, so as to take phrases into account. In the future, we plan to test other approaches for the extraction of phrases, for example by applying algorithms for recognising multiword expressions. We also plan to integrate our best model for phrase simplification in ERNESTA (Barlacchi and Tonelli, 2013), a system for syntactic simplification of Italian documents. Furthermore, within the H2020 SIMPATICO project, we will integrate our phrase simplification approach in the existing services

of Trento Municipality and perform a pilot study with real users.

Acknowledgments

The research leading to this paper was supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819).

References

Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to Stanford: A collection of CoreNLP modules for Italian. *CoRR*, abs/1609.06204.

Gianni Barlacchi and Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, pages 476–487, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA, June. Association for Computational Linguistics.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China, July. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *ACL-IJCNLP 2015 System Demonstrations*, ACL, pages 85–90, Beijing, China.

Gustavo Paetzold and Lucia Specia. 2016a. Benchmarking lexical simplification systems. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Gustavo H. Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3761–3767. AAAI Press.

H. Saggion and G. Hirst. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 347–355, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. pages 901–904.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. SIMPITIKI: a Simplification corpus for Italian. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, volume 1749 of *CEUR Workshop Proceedings*.