Going Dutch: Creating SimpleNLG-NL

Ruud de Jong & Mariët Theune

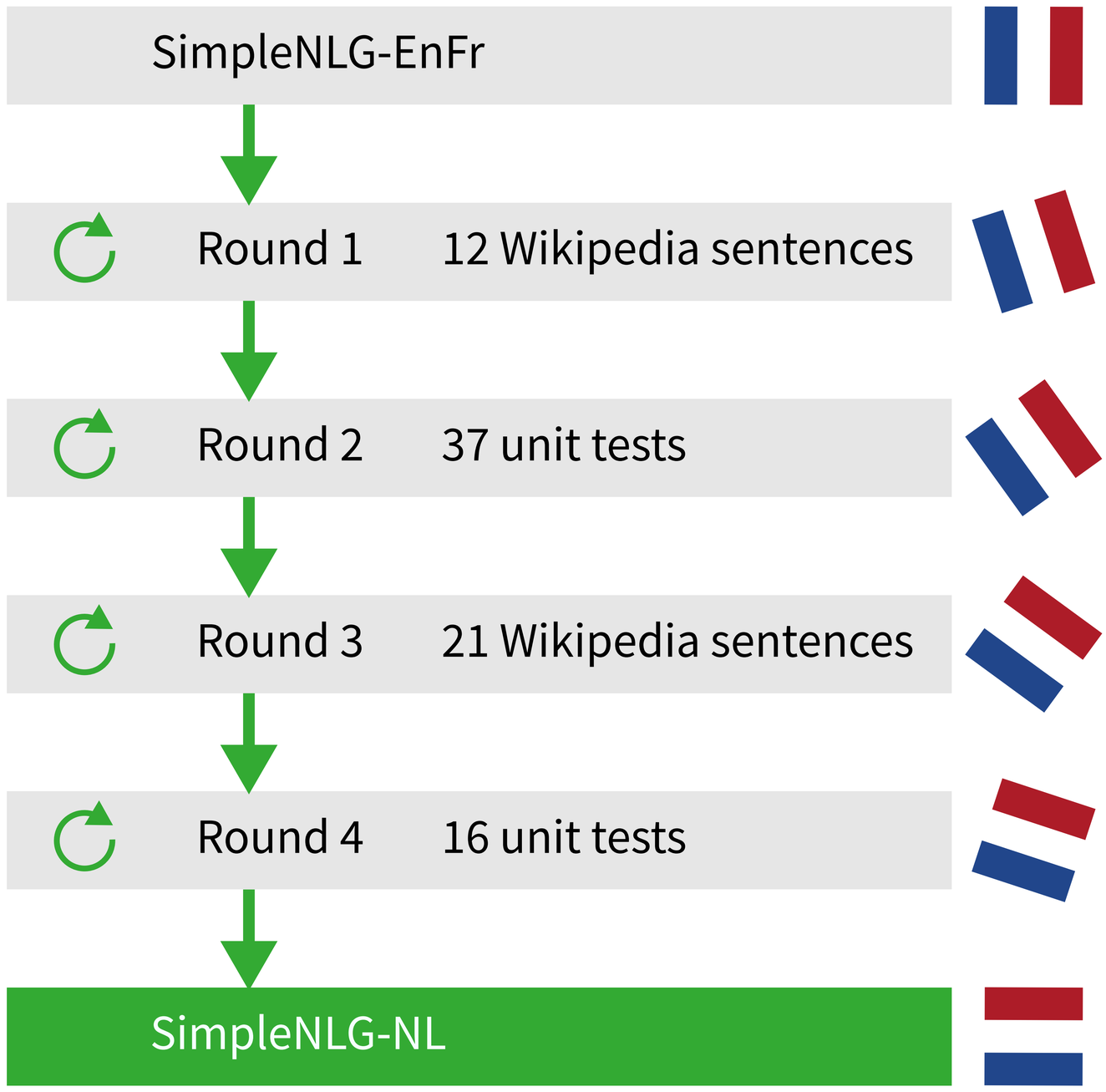
Human Media Interaction, University of Twente, Enschede, The Netherlands

Introduction

We present SimpleNLG-NL, an adaptation of the SimpleNLG surface realisation engine for the Dutch language. It is a Java-based library that can be used to generate Dutch text. The system is part of a family of surface realisers based on the English SimpleNLG. We developed SimpleNLG-NL using a novel method for determining and testing the grammatical constructions to be implemented, using target sentences sampled from a treebank.

Method

SimpleNLG-NL was developed using an iterative generate-evaluate-revise process. As a base, the bilingual SimpleNLG-EnFr was used. The process was divided into four rounds with sentences from Wikipedia. Grammar rules were implemented as needed to generate the target sentences. Rounds 2 and 4 consisted of small unit tests designed to develop and test particular features, such as interrogative sentences.



Results

Sentence set	Sentences	Exact matches		Accepted as correct	
Round 1	12	8	66.7%	11	91.7%
Round 2	37	37	100.0%	37	100.0%
Round 3 (medium)	11	9	81.8%	9	81.8%
Round 3 (long)	10	5	50.0%	7	70.0%
Round 4	16	10	62.5%	10	62.5%
Total	**86**	**69**	**80.2%**	**74**	**86.0%**

SimpleNLG-NL comes with three lexicons based on the Dutch pages on Wiktionary.org. The largest version contains almost 80 000 entries. Smaller versions (selected using a word list of common Dutch words) contain 8600 and 3300 entries, respectively. By default, the medium sized lexicon is used.

Conclusions

Over 80% of the target sentences were generated correctly. We believe that the grammatical structures of the target sentences cover a subset of Dutch grammar large enough to be able to generate simple sentences. With more complex sentences, word order can become incorrect.

Some acceptable differences between target sentences and the generated sentences can be attributed to stylistic choices, which are not supported by SimpleNLG-NL.

The system will be used to realise (parts of) dialogs used in the POSTHCARD project, which uses a simulation of Alzheimer's patients to train caregivers.

SimpleNLG-NL is available at
github.com/rfdj/SimpleNLG-NL
License: MPL 1.1

This research was carried out partially within the POSTHCARD project,
funded by the European AAL programme (aal-call-2017-045).
POSTHCARD is being made possible by ZonMw under project number: 735170004

POSTHCARD
Alzheimer training for caregivers

UNIVERSITY OF TWENTE.