Adapting SimpleNLG for Brazilian Portuguese realisation

Rodrigo de Oliveira
Department of Computing Science
University of Aberdeen
Aberdeen, UK, AB24 3UE
rodrigodeoliveira@abdn.ac.uk

Somayajulu Sripada
Department of Computing Science
University of Aberdeen
Aberdeen, UK, AB24 3UE
yaji.sripada@abdn.ac.uk

Abstract

This paper describes the ongoing implementation and the current coverage of SimpleNLG-BP, an adaptation of SimpleNLG-EnFr (Vaudry and Lapalme, 2013) for Brazilian Portuguese.

1 Introduction

Realisation is the last step in natural language generation (NLG) systems, so the goal of a realisation engine is to output text. SimpleNLG is a Java library that employs morphological, syntactic and orthographical operations on non-linguistic input to output well-formed sentences in English. SimpleNLG-EnFr (Vaudry and Lapalme, 2013) is an adaptation of SimpleNLG for French. This paper describes the current state of SimpleNLG-BP[1], an adaptation of SimpleNLG-EnFr for realisation in Brazilian Portuguese.

2 Recycling SimpleNLG-EnFr

To implement SimpleNLG-BP, we opted to extend SimpleNLG-EnFr instead of the original SimpleNLG. The main reason was the linguistic phenomenon of preposition contraction, which is what happens in *da mesa* (*of the table*): *da* is the fusion of *de* (*of*) with *a* (the.FEM.SNG). Because preposition contraction happens in French but not in English, we simply adapted the algorithm in SimpleNLG-EnFr to suit Brazilian Portuguese.

3 Coverage of SimpleNLG-BP

As of submission date of this paper (May 23, 2014), almost all efforts in implementing SimpleNLG-BP focused on morphological operations, as described in *Moderna Gramática Portuguesa* (Bechara, 2009). However, a testbed of 43 instances including full sentences in non-interrogative form and isolated phrases could be successfully generated by SimpleNLG-BP.

3.1 Morphology

Morphological operations in the current state of SimpleNLG-BP tackle 3 phrase types: noun phrases, preposition phrases and verb phrases.

3.1.1 Pluralisation of nouns

Pluralisation rules in Brazilian Portuguese normally add a final *-s* to nouns, but word-internal modifications may also be applied, depending on the word's stress, last vowel and/or ending. Possible noun endings in Brazilian Portuguese are: *-l*, *-m*, *-n*, *-r*, *-s*, *-x*, *-z* and vowels. SimpleNLG-BP currently includes all pluralisation rules for nouns ending in *-m*, *-r*, *-s*, *-x* or most vowels, but only some rules for endings *-l*, *-n*, *-z* and *-ão*. The pluralisation algorithm will still attempt to pluralise any string, which is useful to handle neologisms.

3.1.2 Preposition contraction

Similar to French, Brazilian Portuguese provides a morphophonological mechanism to contract words in preposition phrases. The prepositions that undergo contraction are *a* (*by*, *to*), *em* (*in*, *or*, *at*), *de* (*from*, *of*) and *por* (*through*, *by*) – or preposition complexes ending in those, such as *atrás de* (*behind*) or *em frente a* (*in front of*). When these precede a determiner or adverb, preposition and following item combine to form a single word. Take *as* (*the*.FEM.PLR), for instance. If it appears in a preposition phrase after *a*, *em*, *de* or *por*, the result will be *às*, *nas*, *das* and *pelas*, respectively. Note that *desde* (*since*) ends with *-de* but does not undergo contraction. The same applies for *contra* (*against*) and *para* (*to*, *for*); both end in *-a* but do not undergo contraction.

[1]The source code for SimpleNLG-BP can be found at https://github.com/rdeoliveira/simplenlg-en-fr-pt .

Proceedings of the 8th International Natural Language Generation Conference, pages 93–94,
Philadelphia, Pennsylvania, 19-21 June 2014.

3.1.3 Verb conjugation

English systematically combines all 3 tenses – past, present and future – to perfective and/or progressive aspects. This gives English a total of 12 possible combinations for the same verb, person and number. Subjunctive or imperative moods are of little concern to English, since base forms of verbs are usually identical to non-indicative forms.

Brazilian Portuguese may be said to express the same number of tenses, aspects and moods. In practice, this does not apply. Perfectiveness in Brazilian Portuguese traditional grammars is seen as a 3-element set – perfective, imperfective and pluperfective – which apply only to the past tense. English uses perfectiveness across all 3 tenses (*had done*, *have done*, *will have done*). Moreover, subjunctive forms in Brazilian Portuguese are morphologically distinct from indicative forms. Conditional is not built by adding an unchangeable auxiliary (e.g. *would*), but by morphology as well. Finally, infinitive forms of verbs may be conjugated or not. Thus, it was more practical to implement tense in SimpleNLG-BP as a 10-element set – past, present, future, imperfect, pluperfect, conditional, subjunctive present, subjunctive imperfect, subjunctive future and personal infinitive – where each tense may already pack some sense of aspect and mood.

Nevertheless, we implement aspect as a separate 3-element set, to be optionally declared as verb features, in order to trigger verb periphrasis formation. Modern Brazilian Portuguese uses verb periphrases extensively; e.g. the periphrastic form *tinha feito* (*had done*) is normally used instead of the single-verb form *fizera* (also *had done*). SimpleNLG-BP associates *ter* (*have*) to perfectiveness and *estar* (*be*) to progressiveness, thereby resembling the grammar of English and preserving most of the optional verb-phrase features used in the original SimpleNLG. Additionally, we included *prospectiveness* in the aspect set (as suggested by Bechara (2009) pp. 214-215) to generate periphrases that express future by means of the auxiliary *ir* (*go*). With a 3-element aspect set and a 10-element tense set, SimpleNLG-BP is able to build 80 different forms[2] for the same verb, person and number. Additionally, negative, passive and modalised verb phrases are also supported. Modals generate prepositions automatically, if required, such as *dar* (be able to) and *acabar* (end), whose prepositions are *para* and *de* respectively.

As far as subject-verb agreement, if the verb to be conjugated exists in the default lexicon file, the final string is simply retrieved; if not, a conjugation algorithm attempts to inflect the verb. For SimpleNLG-BP, we compiled an XML lexicon file out of DELAF_PB (Muniz, 2004), an 880,000-entry lexicon of inflected words in Brazilian Portuguese. The original file became too large at first – 1,029,075 lines, 45.4MB – which turned out to be an issue. A default run of SimpleNLG compiles the default lexicon file *a priori* to store it in memory, so a single run (e.g. 1 test case) took an average of 2.5 seconds, just to build the lexicon onto memory. Since an inefficiency of that dimension can be prohibitive in some practical contexts, we compiled a smaller list of 57 irregular verbs in Brazilian Portuguese plus personal pronouns, which became only 4,075-line long (167KB) and takes only 0.17 seconds for compilation in average. SimpleNLG-BP includes both the lexicon file and the lexicon compiler, if one wishes to modify the default lexicon.

[2]Even though 22 of these don't seem to be used by Brazilian Portuguese speakers.

4 Summary

We described SimpleNLG-BP, an ongoing adaptation of SimpleNLG for Brazilian Portuguese, which currently supports noun pluralisation, preposition contractions and verb conjugation, and includes a lexicon file and a lexicon compiler.

Acknowledgements

The first author of this paper thanks Arria/Data2Text Limited for funding his doctoral research at the University ofAberdeen.

References

Evanildo Bechara. 2009. *Moderna Gramática Portuguesa*. Nova Fronteira & Lucerna, Rio de Janeiro, 37 edition.

Marcelo Caetano Martins Muniz. 2004. A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB. Master's thesis, USP.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting SimpleNLG for bilingual English-French realisation. In *14th European Conference on Natural Language Generation*, pages 183–187, Sofia, Bulgaria.