Towards a Rule Based System for Automatic Simplification of Texts

Jonas Rybing, Christian Smith
Linköping University
jonry526@student.liu.se,
chrsm588@student.liu.se

Annika Silvervarg
Department of Computer and Information Science
Linköping University
annika.silvervarg@liu.se

1 Introduction

The need for simplified texts in various areas is increasing, however manually simplification of texts is very resource intense and costly. An automatic system for simplification of texts is therefore very desirable. In this paper we present the initial development of such a system for Swedish and discuss results from an evaluation based on various mathematical measures for simplified texts.

2 The CogFLUX system

The CogFLUX system is based on transformation rules used to reduce complexity of texts. The rules were compiled by Anna Decker (2003) based on studies of corpora of easy to read texts and normal texts. She has identified 25 general syntactic transformation rules for simplification. The rules can be grouped into two subsets of rules; 1) rules that remove or replace sub phrases and 2) rules that add new syntactical information to the text. An example of a rule from the first category is: np(det+ap+n) → np(n). This rule will replace any nominal phrase containing a determiner, an adjective phrase and a noun with a nominal phrase containing only the noun.

CogFLUX implements the first subset of Decker's rules. The system consist of several modules, eg, GRANSKA tagger for part of speech tagging and Malt parser for syntactic analysis, and a module to replace abbreviations with its extended form.

3 Evaluation measures

The formulas used in this study are Swedish readability index (LIX), noun quota (NQ) and lexical variation (OVIX). These are all mathematical formulas resulting in a strict quantitative value. The advantage of using quantitative measures is that they can be applied automatically and the results are easy to compare.

The measure LIX, developed by Björnsson (1968), has been extensively used to measure the readability of Swedish texts. LIX is calculated using the formula:

$$LIX = \frac{O}{P} + 100\frac{L}{O}$$

where O is the number of words in a text, P is the number of sentences in a text and L is the number of long words, i.e. words with more than 6 characters. Measuring the amount of information in a text can be done with the NQ measure. A result around 100 is regarded as a normal ratio of information representing that of newspapers (Josephson et al., 1990). The information ratio is calculated by:

$$NQ = \frac{Number\ of\ nouns}{Numbers\ of\ Verbs} \times 100$$

OVIX is a ratio measure the number of unique words in the text, representing how rich of a variation of words used in the text. A high value, i.e. rich text, is associated with lower readability (Lundberg & Reichenber, 2008). OVIX is calculated by:

$$OVIX = \frac{Number\ of\ uniqe\ words}{Number\ of\ total\ words}$$

4 Results

The evaluation material used was a collection of texts with a total of 100 000 words, of which 50% was fiction, 25% was newspaper articles and 25% was public authority documents. These were automatically simplified by CogFLUX and the three evaluation measures were then computed for the resulting simplified texts. Seven sets of transformation rules were used in the evaluation. The sets were composed and categorized by what type of phrase the rules manipulated, adjective phrases (AP), noun phrases (NP) or preposition phrases (PP). Some of the sets are combinations of different rules, e.g. a set with rules manipulating both noun and preposition phrases (NP+PP). In table 1 the results of the evaluation are presented. The first column displays the text categories. The second column displays the type of measure and the remaining columns shows which rule set was applied and their resulting value. For comparison, manually written easy to read texts accumulated by Katarina Mühlenbock at the University of Gothenburg is also included in the last column (Manual). The texts in this corpus are distributed accordingly to the distribution of the texts used in this study, but this corpus consisted of about one million words.

The LIX value actually increases slightly for all of the texts regardless of applied rule set. The biggest increase can be found in where the prepositional (PP) and to some extent the noun (NP) phrase rules where applied. When phrases are deleted it will only change

		Rule sets								
		All	NP+ PP	NP+ AP	PP+ AP	PP	AP	NP	No	Manual
Fictive texts	LIX	44	46	42	43	46	42	45	41	24
	NQ	46	46	66	46	46	66	66	66	55
	OVIX	14	15	13	14	15	13	14	14	0,07
Newspaper articles	LIX	56	59	53	56	58	53	56	52	36
	NQ	88	88	126	88	88	126	126	126	123
	OVIX	25	25	22	24	25	21	22	22	32
Authority documents	LIX	54	56	51	51	56	51	54	51	35
	NQ	90	90	122	122	90	122	122	122	90
	OVIX	7	7	6	6	7	6	7	6	1,28
All	LIX	49	52	47	49	51	47	50	46	29
	NQ	64	64	90	64	64	90	90	90	75
	OVIX	12	12	11	12	12	10	11	11	-

Table 1 Evaluation measures for different rule sets and different text genres, as well as manually simplified texts.

the LIX positively if the phrase contained a majority affects the LIX negatively. Another reason for the higher LIX values is that a guideline for easy to read texts, applied in CogFLUX, is to replace abbreviations with the full form. When this occurs a short abbreviation is exchanged with one or more words, long or short, but the total number of sentences remains unchanged, which cause increase in the LIX value.

The measures NQ values drops noticeably when PP rules where applied. In regards to the normal NQ value of 100 the measured 46 for fictional texts is a very low value. This indicates that there is a lower number of nouns, prepositions and participle per word in the texts after the performed simplification.

The OVIX value tends to drop somewhat when the AP rules are applied and increase slightly when the PP rules are used. It therefore seem as adjective phrases contain rarely used words and prepositional phrases contain frequently used words.

The sets of rules used by CogFLUX are manually induced based on ly newspaper articles. There was no observed difference in performance between the different genres. This imply that the rules although induced from one type of texts are general, at least in the aspect of making the same errors and same correct simplifications between the genres.

The values of LIX are considerably lower for the manually generated texts than the values of their automatically generated counterpart. The OVIX values are also lower which can partially be explained by the difference in size of the corpora. The ratio of unique word per words will inevitably drop when a corpus grows bigger. The NQ value is overall lower for the automatically generated texts with the exception of the public authority documents.

5 Discussion

The measurements used should only be seen as indications, with easy to read texts correlating with low values. It was clear that they are not enough to fully determine the readability of a text, as the text often seemed to lose coherence with fragmented sentences despite getting better results on the measures. This indicates that the measurements should be complimented with some way to measure readability on a more grammatical level, the coherence of the whole text, or the relevancy of information kept or deleted.

As of today, CogFLUX accepts all suggestions generated by Decker's transformation rules and performs them accordingly. However, Decker found that there are times when transformation rules not is applicable, thus the transformations should not always be performed. Because of this, the transformations are more often than not performed at the wrong place at the wrong time, effectively deleting important information and resulting in a fragmented text. Thus, the simplification rules are not enough to simplify texts on their own. The system need some way, using decision making or further heuristic, of determining when to apply a rule and when not to.

Reference

C. H. Björnsson. Läsbarhet. Bokförlaget Liber AB, 1968.

A. Decker. Towards automatic grammatical simplification of swedish text. Master's thesis, Stockholm's University, 2003.

O. Josephson, L. Melin, and T. Oliv. Elevtext. Analyser av skoluppsatser från åk 1 till åk 9. Lund: Studentlitteratur, 1990.

I. Lundberg and M. Reichenberg. Vad är lättläst? 2008.