In Proceedings of the 9th Conference of the European Chapter of the ACL (EACL'99), Bergen, Norway

Simplifying Text for Language-Impaired Readers

John Carroll
Guido Minnen
Darren Pearce
Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QH, UK
{johnca,guidomi,darrenp}@cogs.susx.ac.uk

Yvonne Canning
Siobhan Devlin
John Tait
Computing and Information Systems
University of Sunderland
Sunderland SR6 0DD, UK
firstname.lastname@sunderland.ac.uk

1 Introduction

Automatic text simplification for language-impaired readers is a relatively unexplored area in natural language processing. We describe a generic system for text simplification (currently at the prototype stage) incorporating a range of state-of-the-art language processing tools. We are applying the system to help people with aphasia (various language impairments, typically occurring as a result of a stroke or head injury) to understand English newspaper articles[1].

Aphasic people may encounter many problems when reading. It has been demonstrated (Devlin, 1999) that these problems can be of a *lexical* nature since less frequent words are often not readily available, and also of a *syntactic* nature in that particular constructions may pose serious difficulties for understanding. In addition to these general aspects of text, there are also problems specific to newspaper text; for example, the often very compact summary-like first paragraph in an article; long sentences; the use of noun compounds and long sequences of adjectives; and frequent use of the passive. Although there is wide variation in the language problems associated with aphasia, depending on such factors as locus of brain injury, aphasia type, and pre-aphasic literacy level, many aphasic people would benefit from a system of the sort we describe.

We outline below the processing strategy of the system and the user-centered evaluation we intend to carry out. We envisage that the results of this project will be of use not only to aphasic individuals, but also to other groups such as non-native speakers whose comprehension of written English text is restricted by limited foreign language skills.

[1]This work is being carried out on the project 'PSET: Practical Simplification of English Text' funded by the UK EPSRC (refs GR/L53175 and GR/L53175). The first author is supported by an EPSRC Advanced Fellowship. Further information about PSET is available at <http://osiris.sunderland.ac.uk/~pset/welcome.html>.

2 The System

We download the original newspaper articles automatically from the WWW[2], and apply a number of processing stages sequentially.

Lexical Tagger The tagger (Elworthy, 1994) assigns and ranks part-of-speech (PoS) tags for each word in a sentence using a first-order HMM. The tagger includes an unknown word guesser with an accuracy of around 85%, and a large disk-resident lexicon specialised to newspaper text.

Morphological Analyser The morphological analyser (an enhanced version of the GATE project lemmatiser (Cunningham et al., 1996)) is based on finite state techniques, and performs an accurate and efficient inflectional analysis of the words in a text given the PoS assignment made by the tagger.

Parser The parser uses a robust feature-based unification grammar of PoS and punctuation tags (Briscoe and Carroll, 1995), coupled with probabilistic LR disambiguation (Carroll and Briscoe, 1996), assigning the most plausible 'shallow' phrase structure analysis to the PoS network (lattice) returned by the tagger. Coverage of a substantial corpus of general text is around 80%. We will improve coverage by utilising recent grammar learning techniques (Osborne, Submitted) to dynamically improve coverage in a principled and tractable manner.

Anaphor Resolver The anaphor resolution component (the only stage not as yet implemented in any form) will be based on CogNIAC (Baldwin, 1997), but rewritten to take advantage of the preceding processing.

Syntactic Simplifier Aphasic people may have problems with syntactic constructions that deviate from canonical subject-verb-object order.

[2]We are using a local newspaper in the north-east of England, *The Sunderland Echo*, that is also published online.

Thus, passive sentences such as *The scheme was singled out by a recent Government report* are found difficult[3], despite the presence of the syntactic cues *was*, *-ed* and *by*. We therefore replace passive constructions with corresponding active forms. We are currently integrating further rules to split conjoined sentences and extract embedded clauses. Syntactic simplification operates iteratively until a configuration is reached that cannot be simplified. This approach is broadly similar to that proposed by (Chandrasekar et al., 1996).

One of the many challenges in syntactic simplification is the observed effect of the total length of a text being increased when longer sentences are replaced by multiple shorter ones. Also, the removal of cohesive devices such as conjunctions may result in anaphora crossing sentence boundaries. To maintain text coherence and cohesion (Grodzinsky et al., 1993) an anaphor is replaced by its referent if the containing sentence is split.

Lexical Simplifier The lexical simplifier (based on (Devlin, 1999; Devlin and Tait, 1998)) replaces content words with simpler synonyms. It first retrieves a set of synonyms for each word from WordNet (Miller et al., 1993), then, according to the user's desired level of simplification, the original word plus a percentage of the synonym list are looked up in the Oxford Psycholinguistic Database (Quinlan, 1992) for the corresponding Kucera-Francis frequencies. The word with the highest frequency is selected.

Morphological Generator Simplification works on the inflectionally analysed text, so the last stage is morphological generation. The generator is simply an inverted version of the morphological analyser described above. The inversion is performed automatically (Minnen and Carroll, Submitted), so any improvements made to the analyser are reflected in the generator at no extra cost. Finally, inter-word spelling changes (e.g. *a apple* $\rightarrow$ *an apple*), auxiliary reduction, etc. are performed.

3 Evaluation

We will perform an experimental evaluation of the system with the help of aphasic participants who are matched to the extent that none display visually related reading difficulties, which would confound the results, and all possess a sufficiently high reading ability—determined at the time of the experiment by using an aphasia assessment battery. As the system is a general tool aimed at all aphasics, the participants will not be screened for aphasia type. The readability of the simplified text and the usability of the system will be assessed by observation and interview; questions will be posed to gauge subjects' comprehension of both explicit and implicit material.

[3]Semantically reversible sentences such as *The boy was kissed by the girl* are even more difficult, since either noun phrase could be the subject.

References

B. Baldwin. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop sponsored by the ACL (Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts)*, Universidad Nacional de Educacion a Distancia, Madrid, Spain.

E. Briscoe and J. Carroll. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proceedings of the 4th ACL/SIGPARSE International Workshop on Parsing Technologies*, pages 48–58.

J. Carroll and E. Briscoe. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 92–100.

R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*.

H. Cunningham, Y. Wilks, and R. Gaizauskas. 1996. GATE—a general architecture for text engineering. In *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*.

S. Devlin and J. Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. In J. Nerbonne. *Linguistic Databases*. Lecture Notes. Stanford, USA: CSLI Publications.

S. Devlin. 1999. Simplifying natural language text for aphasic readers. Ph.D. Dissertation, University of Sunderland, UK.

D. Elworthy. 1994. Does Baum Welch re-estimation help taggers? In *Proceedings of the 4th ACL Conference on Applied Natural Language Processing*, pages 53–58.

Y. Grodzinsky, K. Wexler, Y. Chien, S. Marakovitz, and J. Solomon. 1993. The breakdown of binding relations. *Brain and Language*, 45(3):396–422.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. 1993. Five papers on WordNet. Technical report, Princeton University, Princeton, N.J.

G. Minnen and J. Carroll. Submitted. Fast and robust morphological generation in a practical NLP system.

M. Osborne. Submitted. Minimum description length-based models for practical grammar induction.

P. Quinlan. 1992. *The Oxford Psycholinguistic Database*. Oxford University Press.