Practical Simplification of English Newspaper Text to Assist Aphasic Readers

John Carroll, Guido Minnen
Cognitive and Computing Sciences, University of Sussex
Falmer, Brighton BN1 9QH, United Kingdom

Yvonne Canning, Siobhan Devlin, John Tait
Computing and Information Systems, University of Sunderland
St. Peter's Campus, Sunderland SR6 0DD, United Kingdom

Abstract

Aphasia is a disability of language processing often suffered by people as a result of a stroke or head injury. In order to assist aphasic readers we are developing a system which automatically simplifies English newspaper texts as available on the Internet. The system combines state-of-the-art natural language processing tools with innovative research on text simplification. We present the architecture of the system, discuss the analysis of newspaper text and a number of criteria for simplification. In addition, we provide some initial implementation details and propose an evaluation method.

Keywords: robust parsing, text simplification, aphasia, reading assistance

Introduction

Recently, there has been increasing interest in the use of results from natural language processing for the development of assistive technology.[1] Here, we address this topic by reporting preliminary work carried out in the research project "PSET: Practical Simplification of English Text".[2] The aim of the PSET project is to develop a system to assist people suffering from *aphasia*[3] (or dysphasia)—language impairment typically as a result of a stroke or head injury—in reading newspaper texts.

Aphasia is a huge problem worldwide: the National Aphasia Association reports that one million Americans have aphasia, and the British charity Action for Dysphasic Adults puts the figure for the UK at 250,000. Though the language impairments of aphasic people can be quite diverse in character, it is likely that a great many aphasic people will at some time encounter some problems in understanding written text. In order to assist aphasic people with respect to this aspect of their impairment, we are developing a system which automatically simplifies English newspaper texts as available on the Internet. The system combines state-of-the-art natural language processing tools with innovative research on text simplification. We present the architecture of the system, discuss the analysis of newspaper text and a number of criteria for simplification. In addition, we provide some initial implementation details and propose an evaluation method.

It is generally argued that the impairments of aphasic people can provide a window on normal language function (see, for example, (Howard & Hatfield 1987) and (Shallice 1988)). We therefore envisage that the research we carry out through the development of this system and its evaluation will not only be of use to aphasic individuals, but might also assist normal, non-native speakers whose access to written English text is restricted by limited foreign language skills.

The organisation of the remainder of the paper is as follows. In the next section, we give an overview of the architecture of the system. We then discuss the syntactic analysis of newspaper text as performed by our system, the reasons why text simplification is necessary, and the techniques we propose to use. Finally we describe the set of field experiments we intend to perform to evaluate the implemented system, and make some concluding remarks.

Overview

Figure 1 gives an overview of the architecture of the system we are developing. The system can roughly be divided into two main components: an *analyser* component which provides a syntactic analysis and (partial) disambiguation of the newspaper text, and a *simplifier* component which subsequently adapts the output of the analyser to aid readability for aphasic people.

The analyser consists of three subcomponents: a lex-

[1] See, for example, the working notes of the 1996 AAAI Fall Symposium on Developing Assistive Technology for People with Disabilities, and reports on the first and second workshop on natural language processing and communication aids (1996 and 1997) at <http://alpha.mic.dundee.ac.uk/~slanger/workshop.html>.

[2] PSET is a three-year project funded by the UK Engineering and Physical Sciences Research Council (refs GR/L50105 and GR/L53175), and Apple Computer Inc. 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>.

[3] For a detailed discussion of the various medical and linguistic aspects of aphasia see, for example, (Albert *et al.* 1981), (Caplan & Hildebrandt 1988), (Lesser 1978) and (Lesser & Milroy 1993).

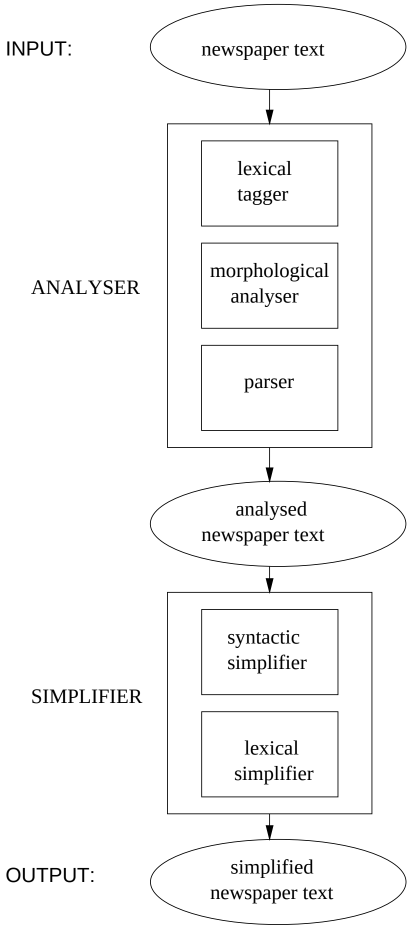

Figure 1: *System architecture*

ical tagger, a morphological analyser and a parser. The simplifier component consists of two subcomponents: a lexical simplifier and a syntactic simplifier. Each of the subcomponents of the system are described briefly in the next sections.

Both the input and the output of the system are intended to be marked-up text documents: the 'markings' themselves are passed through the subcomponents of the system unchanged. Specifically, our input is HyperText Markup Language (HTML) documents, and we fetch these automatically from the online version of the Sunderland Echo newspaper[4] (the local daily newspaper of a city in the north-east of England).

Analyser

The analyser is an enhanced and extended version of a pre-existing system for robust domain-independent syntactic parsing of English, using a unification-based grammar of part-of-speech (PoS) and punctuation labels coupled with a probabilistic LR parser. Currently, the system is able to compute an analysis for around 80% of sentences in a substantial corpus of general text containing a number of distinct genres (Carroll & Briscoe 1996). Many of the 'failures' are due to a requirement to find a root sentence, and not being able to find one in fragments from dialogue, etc. We intend to relax this stipulation, and also make use of recent grammar learning techniques—for example, in (Osborne Submitted)—to dynamically improve coverage in a principled and tractable manner. The system achieves accuracy that is comparable to the state-of-the-art: on a random sample of 250 in-coverage sentences the system has a mean crossing bracket rate of 0.71 and bracket recall and precision of 83% and 84% respectively when evaluated against manually-disambiguated analyses.

Lexical tagger The first subcomponent of the analyser is the lexical tagger, a first-order HMM PoS and punctuation tag disambiguator, which assigns and ranks tags for each word and punctuation token in sequences of sentences (Elworthy 1994). The tagger includes a recently-developed unknown word guesser with an accuracy of around 85%; however we are currently devoting effort to creating a customised, large lexicon that covers a larger proportion of words in newspaper text, together with an efficient disc-based access mechanism.

Morphological Analyser The morphological analyser is a robust, efficient tool based on finite state techniques that performs an inflectional analysis of the words in a text, given the PoS assignment made by the tagger. This component is an enhanced version of the GATE project lemmatiser (Cunningham, Wilks, & Gaizauskas 1996).

Parser The parser uses a feature-based unification grammar of PoS and punctuation labels, assigning 'shallow' phrase structure analyses to tag networks (or 'lattices') returned by the tagger (Briscoe & Carroll 1995). Off-line compilation of the grammar is used to improve run-time parsing efficiency (Carroll 1994): the parser is able to construct a representation of the full set of parses for a sentence in time that is empirically only quadratic in sentence length. The parser uses probabilistic information acquired from training on treebanked corpora, and returns ranked analyses (Carroll & Briscoe 1992; Briscoe & Carroll 1993). As an example, the structure output by the analyser for the (headline) sentence from the Sunderland Echo *Elderly Couple Hurt in Gas Blast* is given in figure 2.

Simplifier

Aphasic people may encounter many problems when reading. (Devlin Forthcoming) shows that these problems can be of a *lexical* nature as less frequent words are often not readily available, and also of a *syntactic*

[4] The online version of the newspaper is accessible through <http://www.sunderland.com/echo/index.htm>.

```
(S (N1 (AP ("elderly" JJ))
       ("couple" NN1))
   (VP ("hurt" VV0)
       (PP ("in" II)
           (N1 ("gas" NN1)
               ("blast" NN1)))))
```

Figure 2: *Example output of the analyser for a (headline) sentence*

nature in that particular constructions may pose serious difficulties for understanding. In addition to these general aspects of text that constitute problems for aphasic readers, there are also problems caused specifically by newspaper text such as the typical very compact summary-like first paragraph in an article. Sentence length can be problematic for aphasic people, and broadsheet newspapers tend to include sentences of around 32–35 words. Even in local papers the average is 16–20 words (Keeble 1994). A common feature of 'tabloidese' is the frequent use of compounds and adjectives, for instance in phrases such as *Twenty-five-year-old blonde-haired mother-of-two Jane Smith.* Although in-house newspaper style books warn against overuse of the passive, newspapers want to attract the attention of the potential reader and so they employ more sensational sentences like *A bid to build an incinerator on local wasteland was today accepted by the council* rather than the more straightforward (and easier for aphasic readers) *The council today accepted a bid to build an incinerator on local wasteland.*

Syntactic simplifier Aphasic people have problems with syntactic constructions that deviate from canonical Subject-Verb-Object order. For example, passive sentences like *The boy was kissed by the girl* where interchanging the subject and the object results in a semantically acceptable sentence, are problematic. Despite the presence of the syntactic cues *was*, *-ed* and *by*, aphasic readers have difficulty understanding such a sentence. In order to assist aphasic readers, we therefore propose to replace passive constructions with active constructions.

Other syntactic simplifications that significantly improve the readability of newspaper text include the elimination of multiply embedded prepositional and relative phrases, and generally the replacement of longer sentences with two or more shorter ones.

We are in the process of building a rule based syntactic simplifier which will convert passives into actives, split conjoined sentences and extract embedded clauses. The simplifier will utilise unification pattern matching over phrase marker trees produced by the analyser, and will operate iteratively, repeatedly applying rules to simplified trees until forms are reached that cannot be simplified further. It is expected rule sets will be ordered and that possibly single-shot rules may be required although the feasibility of this is unclear at this stage. This approach is broadly similar to that proposed by (Chandrasekar, Doran, & Srinivas 1996).

There are a number of possible problems and challenges here: for example the maintenance of text coherence and cohesion—for the aphasic as opposed to the normal reader; also the observed effect of the total length of a text being increased when longer sentences are replaced by multiple shorter ones.

Lexical simplifier The lexical simplifier builds on work reported in (Devlin & Tait 1998) and (Devlin Forthcoming). Devlin's simplifier feeds the content words, one at a time, from the analysed newspaper text into the WordNet lexical database (Miller *et al.* 1993). For each word a file is created containing the synonyms held in WordNet for that entry. The simplifier reads the file and extracts a percentage of the synonyms—specified by the user depending upon what level of simplification is required—and interrogates the Oxford Psycholinguistic Database (Quinlan 1992) for the Kucera-Francis frequency of each synonym and the original word. The most appropriate word (with the highest frequency) is selected and written to an output file so reconstituting the newspaper text.

Many words used in isolation are ambiguous. Syntactic analysis can often resolve this ambiguity by considering words as part of a larger construction, especially when an analyser incorporates statistical information. However, sometimes disambiguation is not possible without a deep semantic analysis which is very costly and would thus seriously compromise the practicality of the system we are developing. Since we refrain from an elaborate analysis of the meaning of the text, lexical simplification could possibly change its meaning. However, we believe that in practice this will not turn out to be a problem given the observation that less frequent words—which are thus candidates for simplification—often have a very specific meaning, i.e. are less likely to be ambiguous.

Evaluation

We will perform an experimental evaluation of the system in the field using as subjects aphasic people who do not have reading problems related to vision.[5] Furthermore, we will ensure that we test only people who possess a sufficient level of comprehension, by restricting the experiments to people who score between 3 and 8 on the sentence reading (comprehension) subtest of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan 1983).

For the purpose of the experiments, the newspaper text will be presented to each aphasic subject using

[5] Evaluating our system with aphasic people who also suffer from visually related reading problems would confound our experimental results.

a laptop computer.[6] The experiments are planned to take place in a closely monitored and supported setting probably within the aphasic person's own home. Assessment of the readability of the simplified text and the usability of the system will be made by observation and interview.

Conclusion

We have described a system to assist aphasic readers that we are currently developing that combines state-of-the-art natural language processing tools with innovative research on text simplification. After an automatic linguistic analysis and (partial) disambiguation of newspaper text, the system applies both syntactic and lexical simplification techniques to improve the readability of the text. The system will be evaluated using reading experiments with aphasic people. We envisage that the results of this project will not only be of use to aphasic individuals, but might also assist normal, non-native speakers whose access to written English text is restricted by limited foreign language skills.

Acknowledgements The authors wish to thank Branimir Boguraev, Mark Hepple, Adam Kilgarriff, Ruth Lesser and Nicolas Nicolov for valuable comments and discussion. Of course, the authors are responsible for any errors.

References

Albert, M.; Goodglass, H.; Helm, N.; Rubens, A.; and Alexander, M. 1981. *Clinical Aspects of Dysphasia.* New York: Springer.

Briscoe, E., and Carroll, J. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics* 19(1):25–60.

Briscoe, E., and Carroll, J. 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*, 48–58.

Caplan, D., and Hildebrandt, N. 1988. *Disorders of Syntactic Comprehension.* MIT Press.

Carroll, J., and Briscoe, E. 1992. Probabilistic normalization and unpacking of packed forests for unification-based grammars. In *Proceedings of AAAI Fall Symposium on Probalistic Approaches to Natural Language. 1992*, 33–38.

Carroll, J., and Briscoe, E. 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*, 92–100.

Carroll, J. 1994. Relating complexity to practical performance in parsing with wide-coverage unification grammars. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics (ACL-94)*, 287–294.

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

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

Devlin, S., and Tait, J. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. In (Nerbonne 1998).

Devlin, S. Forthcoming. *Simplifying Natural Language Text for Aphasic Readers.* Ph.D. Dissertation, University of Sunderland, UK.

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

Goodglass, H., and Kaplan, E. 1983. *The Assessment of Aphasia and Related Disorders.* Philadelphia: Lee and Febiger.

Howard, D., and Hatfield, F. 1987. *Aphasia Therapy.* London: Lawrence Earlbaum Associates Ltd.

Keeble, R. 1994. *The Newspaper Handbook.* Routledge, London.

Lesser, R., and Milroy, L. 1993. *Linguistics and Aphasia: Psycholinguistic and Pragmatic Aspects of Intervention.* Longman.

Lesser. 1978. *Linguistic Investigations of Aphasia.* Edward Arnold Publishers Ltd.

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K.; and Tengi, R. 1993. *Five Papers on WordNet.* Princeton University, Princeton, N.J.

Nerbonne, J., ed. 1998. *Linguistic Databases.* Lecture Notes. Stanford, USA: CSLI Publications.

Osborne, M. Submitted. Minimum description length-based models for practical grammar induction. Submitted to Machine Learning.

Petheram, B. 1988. Enabling stroke victims to interact with a microcomputer: a comparison of input devices. *International Disability Studies* 10:73–83.

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

Shallice, T. 1988. *From Neuropsychology to Mental Structure.* Cambridge University Press.

Most of the referenced papers written by the authors are available via the PSET project home page at <http://osiris.sunderland.ac.uk/˜pset/welcome.html>

[6] (Petheram 1988) provides evidence that even elderly aphasic people respond positively to the use of computers.