The CW Corpus

Matthew Shardlow

1 BASIC INFORMATION

1.1 Corpus Composition

The corpus consists of 731 sentences, mined from Simple Wikipedia edit histories.

1.2 Representation of the corpora

The corpus is available as a text document.

1.3 Character encoding

The characters are encoded in UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact Person

Name: Matthew Shardlow
Address: IT301, IT Building, School of Computer Science, Manchester, M13 9PL
Affiliation: School of Computer Science, University of Manchester
Position: Research Student
Telephone: N/A
Fax: N/A
e-mail: m.shardlow@cs.man.ac.uk

2.2 Delivery Medium

The resource is available on the MetaShare platform as an archive.

2.3 Copyright Statement

The resource is freely available for research purposes.

3 TECHNICAL INFORMATION

3.1 Directories and Files

The corpus is available as a text file (.txt) which contains one entry per line.

3.2 Data Structure of an Entry

Each entry consists of 4 tab separated fields in the form: Sentence Index, Complex Word (CW) Index, Suggested Replacement, Sentence.

Sentence Index: A number in the form ### which uniquely identifies each sentence.

CW Index: An Integer which locates the complex word in the sentence. Counted from 0 for computational simplicity. (i.e. if the index is 3, then the CW will be the 4th token in the sentence.)

Suggested Replacement: The replacement suggested by a Simple Wikipedia editor.

Sentence: The sentence boundaries are indicated by square brackets. Tokens are separated by spaces.

3.3 Corpora Size

There are 32590 tokens in the corpus and it is 174KB in size.

4 CONTENT INFORMATION

4.1 Type of the corpus (4.1 (monolingual/multilingual, parallel/comparable, raw/annotated)

This is a monolingual, annotated corpus.

4.2 The natural language(s) of the corpus

The language of this corpus is English.

4.3 Domain(s)/register(s) of the corpus

The corpus contains sentences mined from Simple Wikipedia edit histories.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Lexical mark-up. One CW per sentence is identified. We take a broad view on the definition of a CW, which may be loosely summed up as: any word which reduces the readability and understandability of it's containing sentence. This measure of complexity may change from sentence to sentence and from reader to reader. The typical readers that these simplifications are aimed at are the same as those of Simple Wikipedia: people with low English language proficiency.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc tagged or parsed)

N/A

4.4.3 Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

N/A

4.4.4 Attributes and their values (if annotated)

N/A

4.5 Intended application of the corpus

The evaluation of CW identification systems.

4.6 4.6 Reliability of the annotations (automatically/manually assigned) if any

The reliability was assessed by 6 annotators who gave a combined accuracy score of 90.67% with a Fleiss' Kappa of 0.68, rising to 97.5% with a Kappa of 1 if the 2 lowest agreeing annotators are removed.

RELEVANT REFERENCES AND OTHER INFORMATION

[1] Matthew Shardlow, *The CW Corpus: A New Resource for Evaluating the Identification of Complex Words.* In Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations Addison Wesley, Massachusetts, at the 51st Annual Meeting of the

Association for Computational Linguistics, Sofia, Bulgaria, Association for Computational Linguistics.

[2] Matthew Shardlow, *A Comparison of Techniques to Automatically Identify Complex Words*. In Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, Association for Computational Linguistics