RC23994 (W0606-147) June 30, 2006
Computer Science

IBM Research Report

EasyEnglishAnalyzer: Taking Controlled Language from Sentence to Discourse Level

Arendse Bernth
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

EasyEnglishAnalyzer: Taking Controlled Language from Sentence to Discourse Level

Arendse Bernth
IBM T.J. Watson Research Center
19 Skyline Dr.
Hawthorne, NY 10532
arendse@us.ibm.com

Abstract

Controlled Language checking has traditionally been applied largely on a sentence level by placing restrictions on permissible vocabulary and permissible syntactic constructions, including proper punctuation. Only little attention has been paid to document or discourse level checking. In this paper, we report on work on EasyEnglishAnalyzer to handle certain discourse and document level checks. This work helps take Controlled Language checkers from the sentence level to the discourse and document level.

Deep semantic analysis provided by a discourse understanding system assists with the semantically more challenging tasks such as proper paragraph structure. For checks related to overall style and organization, document structure is recognized by enhanced interpretation and use of document structure tags so that appropriate checks can be applied in a context-sensitive manner. Parsing of combined segments is applied in checking correctness of list environments.

1 Introduction

Traditionally, controlling of language has been applied at the sentence level by placing restrictions and recommendations on vocabulary and syntactic constructions. In this paper, we report on work taking Controlled Language (CL) checkers to the next level, that of the text and document.

According to [O'Brien, 2003], CL rules can be classified in the following way:

1. Lexical rules
2. Syntactic rules
3. Textual rules:
   (a) Text Structure
   (b) Pragmatic

In order to compare CL rule sets across a number of CLs, [O'Brien, 2003] goes on to list the linguistic subcategories for each of the above main categories. The categories of interest for the current paper are the subcategories pertaining to *Text Structure* and *Pragmatic*.

Most of the rules in these subcategories are sentence-internal, such as sentence length, punctuation and verb form usage, with the exception of paragraph length and structure, and to our knowledge these have not been implemented in a CL checker, but are rules to be followed by the writers without the assistance of a checking tool. AECMA Simplified English is by far the most "text-aware" of the CLs being compared [O'Brien, 2006], but even for Boeing's CL checker, the only check with a wider scope than the sentence is a count of the sentences in a paragraph.[1] Generally speaking, with the exception of AECMA Simplified English, which does exhibit some textual awareness, CL rules apply within the sentence.[2]

In this paper, we report on work taking EasyEnglishAnalyzer (EEA) [Bernth, 1997, Bernth, 1998b, Bernth, 2000] from the sentence level to the discourse and document level. EEA is an authoring tool that helps writers produce clearer and less ambiguous English. Where appropriate, EEA provides rephrasing suggestions. Like other CL checkers, EEA was sentence-based. However, recently requirements came up for checks that in effect take EEA from the sentence level to the document level. These requirements range from the mundane (but still complicated), such as checks for proper punctuation in bulleted lists, to the very ambitious, such as "one thought per paragraph". The requirements reflect, in most cases, traditional good writing style such as described in e.g. [Elsbree and Bracher, 1967, Glorfeld *et al.*, 1974, Strunk, 2000]

EEA depends on an English Slot Grammar (ESG) [McCord, 1980, McCord, 1990] parse expressed as a network, as described in [Bernth, 1998a]. Disambiguation rules explore the network to spot ambiguous or potentially ambiguous structures. Syntactic rules explore the network to spot syntactically infelicitous structures. Lexical rules explore the network to check conformance to vocabulary restrictions, some of which depend on contextual information provided by the parse. Like other parsers, ESG is sentence-based. The challenge was then to expand the existing technology, ESG and EEA, to be able to handle checks that apply to text spanning several sentences, such as paragraphs and the whole document. Part of this task involved some rather uninteresting, but nevertheless non-trivial, engineering relating to memory management etc, which we will not discuss further in this paper.

ESG and EEA already make use of document structure tags, so it was obvious to expand the use of these to help get the document structure, which is important because the decision on which checks to apply depends on which part of the document is being checked. So we needed to set up an infrastructure that keeps track of what type of text EEA is processing, and change the checks applied accordingly. Additionally, it was decided to integrate the discourse understanding system Euphoria [Bernth, 2002, Bernth, 2004] into EEA to assist with the more semantically demanding checks.

Euphoria produces a semantic analysis spanning several sentences (a *discourse*) with coreference resolved and implicit arguments made explicit. The semantic analysis uses entity-

[1] Richard Wojcik, personal communication June 27, 2006.
[2] Sharon O'Brien, personal communication June 21, 2006.

oriented logical forms, built around a notion of extended entities – generalized event variables – and covers the complete text, but is neatly indexed by entity variables.

The paper is organized as follows. Section 2 describes requirements related to general writing style and document organization, and the handling of these requirements. In Section 3 we take a closer look at bulleted lists. Section 4 reports on paragraph level control, such as paragraph topics, and Section 5 gives our conclusion.

2 Style and organization

There were a number of requirements relating to writing style and document organization, two of which we report on here. Section 2.1 describes the requirement for general compelling writing style, which we treated as a requirement for sentence variation, and Section 2.2 briefly describes text size and organization.

2.1 Style

The most approachable requirement for style was for good "rhythm", which was treated as a requirement for proper sentence variation. Sentences may vary in voice, sentence length, sentence type (declarative, interrogative, imperative, or noun phrase), and running text versus bulleted text [Glorfeld *et al.*, 1974], p.39. We were able to get this directly out of the parses and the document structure tags, with added infrastructure to keep track of the properties across sentences. Passive voice and questions are, generally speaking, to be minimal, so some judgement about the proper proportion of these relative to "normal" segments is called for.

Figure 1 [IBM, 2006b] illustrates good rhythm; this is particularly noticeable if you read the text aloud. Initially, three noun phrases set up a certain rhythm, which is then broken by a question of normal sentence length at the same time reflecting the rhythm of three set up by the previous segments. This is followed by a longer declarative sentence. Finally, the paragraph is wrapped up by another occurrence of three short segments, this time imperative segments. As the dot on the *i*, a one-word segment, an adverb, that applies to the three imperatives.

Drug pipelines under pressure. Rising development costs. Longer submissions processes. How do you get new drugs to market faster, reduce costs, and increase business value? IBM provides pharmaceutical companies with integrated, flexible, efficient infrastructure solutions that enable deeper analysis, closer collaboration. Find more leads. Automate supply chain management. Smooth FDA submissions. Faster.

Figure 1: Sentence variation.

Even though sentence variation was generally called for, certain types of sections are required to be written in purely narrative format: no questions, no fragments, no bullets, etc.

2.2 Organization and text size

In order for the text to have an easy-to-read appearance as well as fit properly into the web page template that the writers are required to use, the text has to be organized in smaller, pre-defined sections of pre-defined length, and paragraphs should not be overly long. Handling this requirement depends heavily on identifying document structure, and once this is accomplished the task is mainly a matter of counting characters.

Counting characters is not *quite* as trivial as it might seem at first blush since the input text is in HTML and the characters to count are the characters displayed on the web page. HTML entities taking up several characters in the input may take up only one character when formatted, and hence need to be recognized and the count adjusted accordingly.

3 Bulleted lists

For bulleted lists we need to consider the sentence leading in to the list (the "lead-in") as well as the list itself, and this combination has its own set of requirements, some of which were at odds with earlier requirements designed to improve the quality of sentence-based Machine Translation (MT) [Bernth and Gdaniec, 2001].

For MT purposes, it is important that each list element is a complete segment that can stand on its own, since this is the way it will be treated by most MT systems. If, for example, the text introducing the list is complemented by a bare infinitive at the beginning of each list element, then a sentence-based MT system may interpret the infinitive as an imperative, with disastrous results for translation into languages with a distinction between infinitives and imperatives.

Traditional stylebooks frown on this practice independently of MT. [Elsbree and Bracher, 1967], p. 475, says (our emphasis):

> Note that a list introduced by a colon should be in apposition to a preceding word; that is, *the sentence preceding the colon should be grammatically complete without the list.*

And [Strunk, 2000], p. 8, says:

> It [a colon] usually follows an independent clause and should not separate a verb from its complement or a preposition from its object.

However, in order to make the text more catchy, exactly this structure was considered desirable as illustrated in Figure 2 [IBM, 2006a]. This structure is supposed to draw in the reader to continue reading.

List elements can be of various types, such as complements of the lead-in sentence, complete sentences, or noun phrases. A requirement was that they be of the same type within the same list – parallellism is required.

To support your intense computing needs, IBM Linux Clusters help you:

- *Process data more quickly to better understand molecular behaviors, and construct and refine molecular models on the fly.*
- *Leverage the value of open source operating systems to power Tripos applications.*

Figure 2: Bulleted lists.

The type of list element furthermore determines the proper terminal punctuation both for lead-in and list elements. If the list element is a phrase that continues the lead-in sentence, then the lead-in should end with a colon; if the lead-in is a complete sentence, then it should be terminated with a period. The EEA process is a one-pass process, so we need to remember continually the currently last sentence, which *could* be the lead-in sentence to a list. When we get to a list start markup tag, we then compare the last sentence and its terminal punctuation with the first list element and look for possible mismatches.

Mismatches can be of two kinds. One mismatch is in the terminal punctuation of the lead-in; the other is in the terminal punctuation of the list elements. In order to decide whether the list item continues the lead-in we do two things. We check if the lead-in has missing final complements, and we check if the lead-in combined with the list element constitutes a parsable segment.

If the lead-in and the list elements combine to parsable segments, we also know that the list elements are not to be considered complete sentences. This affects the terminal punctuation of the list elements because elements must end in a period *only* if they are complete sentences.

A requirement that we have not addressed yet, but which will draw heavily on Euphoria is that list elements should be ordered so there is a logical flow. This obviously is a very ambitious requirement that necessitates deep semantic analysis. Using Euphoria to identify sentence and paragraph topic as touched on in Section 4, in conjunction with domain knowledge, is a high-level view of the planned approach for handling this requirement.

4 Paragraph topic

The requirements for paragraphs were traditional recommendations for good paragraphs. Ideally, each paragraph should have exactly one topic. The topic of the first sentence of the paragraph should introduce the overall topic of the paragraph on a high level, acting as a *topic sentence* [Strunk, 2000], p. 16; [Glorfeld *et al.*, 1974], pp. 123, 127. There should also be a relationship between the topic sentence and the nearest title.

An example of such a paragraph structure is given in Figure 3 [IBM, 2006a]. The header ***Focus on expanding your drug pipeline*** sets the overall theme. The topic sentence *Move more winning compounds into clinical trial faster.* presents the main way of doing this, and the rest of the paragraph elaborates on this point.

The coreference aspect of Euphoria is used as a preliminary (and crude) way of detecting

Focus on expanding your drug pipeline
Move more winning compounds into clinical trial faster. With the cheminformatics solution from IBM and Tripos, you can build a virtual discovery lab to quickly test and better understand molecular behaviors. Search and screen disparate molecular databases. Find compounds matching pharmacophore or receptor-site constraints, and dock conformationally flexible ligands.

Figure 3: Good paragraph structure.

new topics. If a non-initial sentence does not have any referents that appeared earlier in the paragraph, then there's obviously a new topic that needs to be flagged.

A better approach, which is not fully developed yet, is to use Euphoria to identify the topic of a sentence. A comparison of the topic of the topic sentence (see above) with the individual topics for the following sentences can then be used to reveal any plurality of topics.

Another requirement relevant for paragraph level checking is that of sentence variation, as described above in Section 2.1.

5 Conclusion

Expanding the range of CL checking from the sentence level to the document level poses a lot of interesting challenges, which are far from exhausted. Functionality accomplished so far needs to be refined and enhanced, and further areas addressed. Our results so far, even though preliminary, indicate the possibility of CL checking on a level that has up till now been more or less ignored, and we invite others in the field to join in this new and exciting aspect.

References

[Bernth and Gdaniec, 2001] Arendse Bernth and Claudia Gdaniec. MTranslatability. *Machine Translation*, 16:175–218, 2001.

[Bernth, 1997] Arendse Bernth. EasyEnglish: A tool for improving document quality. In *Fifth Conference on Applied Natural Language Processing*, pages 159–165, Washington, DC, USA, 1997. Association for Computational Linguistics.

[Bernth, 1998a] Arendse Bernth. EasyEnglish: Addressing structural ambiguity. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Machine Translation and the Information Soup, Third Conference of the Association for Machine Translation in the Americas*, number 1529 in Lecture Notes in Artificial Intelligence, pages 164–173, Langhorne, PA, USA, 1998. Association for Machine Translation in the Americas, Springer.

[Bernth, 1998b] Arendse Bernth. EasyEnglish: Preprocessing for MT. In *Proceedings of the Second International Workshop on Controlled Language Applications, (CLAW-98)*, pages 30–41, Pittsburgh, PA, 1998. Carnegie-Mellon University.

[Bernth, 2000] Arendse Bernth. EasyEnglish: Grammar checking for non-native speakers. In *Proceedings of the Third International Workshop on Controlled Language Applications*, pages 33–42, Seattle, WA, 2000.

[Bernth, 2002] Arendse Bernth. Euphoria – A reference resolution system for machine translation. Technical Report RC22627, IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, November 2002.

[Bernth, 2004] Arendse Bernth. Euphoria semantic analysis. Technical Report RC23396, IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, November 2004.

[Elsbree and Bracher, 1967] Langdon Elsbree and Frederick Bracher. *Heath's College Handbook of Composition*. D.C. Heath and Company, Boston, MA, 1967.

[Glorfeld *et al.*, 1974] Louis E. Glorfeld, David A. Lauerman, and Norman C. Stageberg. *A Concise Guide for Writers*. Holt, Rinehart and Winston,Inc., New York, NY, 1974.

[IBM, 2006a] IBM. http://www-03.ibm.com/industries/healthcare/doc/content/solution/996901105.html, 2006.

[IBM, 2006b] IBM. http://www-03.ibm.com/industries/healthcare/doc/jsp/indseg/ pharmaceutical/index.jsp, 2006.

[McCord, 1980] Michael C. McCord. Slot Grammars. *Computational Linguistics*, 6:31–43, 1980.

[McCord, 1990] Michael C. McCord. Slot Grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: International Scientific Symposium, Lecture Notes in Computer Science*, pages 118–145. Springer Verlag, Berlin, 1990.

[O'Brien, 2003] Sharon O'Brien. Controlling Controlled English. An analysis of several controlled language rule sets. In *Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, (EAMT-CLAW 03)*, pages 105–114, Dublin City University, Ireland, 2003.

[O'Brien, 2006] Sharon O'Brien. *Machine Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*. PhD thesis, Dublin City University, Ireland, 2006.

[Strunk, 2000] William Strunk. *The Elements of Style*. Allyn and Bacon, Boston, London, Sydney, Tokyo, Singapore, 2000.