Controlled Languages for Machine Translation: State of the Art

Hiroyuki Kaji
Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo, 185-8601, Japan
e-mail: kaji@crl.hitachi.co.jp

1 Controlled language and related concepts

A *controlled language* is a subset of a natural language with artificially restricted vocabulary, grammar, and style. Texts written in a controlled language are usually less complex and less ambiguous than those written in an uncontrolled language. The use of a controlled language therefore produces better results in machine translation. On the other hand, a controlled language reduces the power of expression and decreases the writing speed. In short, a controlled language brings out the maximum performance of machine translation systems at the cost of the burden on source text authors. So the controlled language approach is suitable for translation for dissemination of information. And a controlled language becomes more beneficial when texts are translated into multiple target languages.

We should note the distinction between a controlled language and a *sublanguage,* which are sometimes confused. The term ‘sublanguage’, which means literally a subset of a language, is used when focus is put on a language used in a specific domain (for example, weather forecasting) rather than on the whole of a language. ‘Sublanguage’ does not imply artificially imposed restrictions. We should also mention *pre-editing.* Pre-editing is a form of human assistance in machine translation. It includes not only rewriting a source text but also inserting special symbols or tags within the text. Pre-editing is not always done by the authors of source texts, but the controlled language is originally expected to be used by the authors themselves.

2 Historical overview of controlled languages

The notion of controlled language originates in Ogden’s Basic English in the 1930s [1]. Basic English was proposed both as an international language and as a foundation for English learning. It consists of 850 basic words, and a number of inflection and derivation rules.

The first practical controlled language was Caterpillar Fundamental English (CFE) [2]. CFE was designed by Caterpillar Inc. from the mid-60s to the 70s so that they could write product documents that are easily understood by non-native speakers of English. It consists of a specialized vocabulary of about 850 words and an extremely limited grammar. CFE inspired a number of controlled languages including Smart’s Plain English Program (PEP) and White’s International Language of Services and Maintenance (ILSAM). Perkins Engines Ltd. reported a successful application of a controlled language to machine translation [3]. That is, Perkins Approved Clear English (PACE) made the rate of post-editing three to four times faster than usual. These and other controlled languages in the 80s are surveyed in [4].

At present the most widely used controlled language is Simplified English (SE) specified by AECMA (European Association of Aerospace Industries). AECMA SE is used as a world-wide documentation standard in the aircraft industry. It is a human-oriented controlled language. That is, it aims at enhancing the readability and consistency of aircraft maintenance documents. The SE Guide, whose first version was released in 1986, prescribes a basic vocabulary of about 3,100 words and a set of 57 writing rules [5].

Development of MT (machine translation)-oriented controlled languages has been carried on by a growing number of organizations. Caterpillar Inc. has replaced CFE with Caterpillar Technical English (CTE) in conjunction with the KANT machine translation system developed by Carnegie Mellon University [6]. CTE has an extended vocabulary of more than 70,000 words, while the vocabulary of CFE was restricted to less than 1,000 words. Other MT-oriented controlled languages include Controlled Automotive Services Language (CASL) at General Motors [7], ScaniaSwedish for truck maintenance documents at Scania, a Swedish truck manufacturer [8], Controlled English at Alcatel Telecom of Belgium [9], and EasyEnglish Language at IBM corporation [10]. Unlike AECMA SE, most of these controlled languages are being used as proprietary by the organization which developed them.

Most existing controlled languages are intended for technical domains, although recently, controlled language approaches to new domains such as WWW product catalogues has been also reported [11]. The type of texts to which controlled languages have been most effectively applied is procedural texts such as operation manuals and maintenance manuals. Application of controlled languages to descriptive texts, which convey various types of information, is harder than that to procedural texts.

3 Contents of controlled languages

A controlled language can be defined by a set of restrictions on vocabulary, grammar, and style. Restrictions necessary or effective for MT-oriented controlled languages do not necessarily coincide to those for human-oriented controlled languages, and vice versa. Typical restrictions are shown below together with comments on their relevance to machine translation.

3.1 Restrictions on vocabulary

Although a restricted vocabulary is fundamental for any controlled language, the significance of the vocabulary size differs between MT-oriented and human-oriented controlled languages. The vocabulary size in MT-oriented controlled languages need not be small, while it should usually be minimized in human-oriented controlled languages. This is because computers are capable of memorizing a large number of words.

The ambiguity of natural language is the biggest problem in machine translation. The most important point for MT-oriented controlled languages is therefore to restrict parts of speech and meanings of each approved word. Restriction on parts of speech reduces syntactic ambiguity of a sentence, and restriction on meanings makes it easy to select target words. In addition to specifying approved/unapproved words or meanings, it is desirable to suggest alternative approved words for unapproved words or meanings. Examples are given below:

- right [adj]
  Approved meaning: on the side of a body that does not contain the heart
  Unapproved meaning: correct
  Alternative approved word: correct
- since [conj]
  Approved meaning: after the time when
  Unapproved meaning: as it is a fact that
  Alternative approved word: because

3.2 Restrictions on grammar

- *Do not use sentences longer than 20 words.*

Restricting the length of a sentence is simple but effective for machine translation, because shorter sentences contain less syntactic ambiguities.

- *Do not make sequences of more than four nouns.*

This restriction eases the difficulty in analyzing the semantic relations between the constituent nouns.

- *Do not use sentences where the governor for a dependent is not the nearest word that can syntactically govern the dependent.*

Although this restriction may be too strict for authors, the effectiveness for machine translation is obvious. In a language with flexible word order like Japanese, a sentence can sometimes be rewritten to one satisfying the restriction. An example is given below:

> “ファイルに (file *ni)* 修正した (modified) 文書を (document *wo)* 保存する (save)" can be re-written to "修正した (modified) 文書を (document *wo)* ファイルに (file *ni)* 保存する (save)”.

Human-oriented controlled languages do not need this type of restriction. This is mainly due to the difference between the amounts of semantic and common-sense knowledge of humans and computers.

- *Do not use coordination that requires distributive reading.*

This is also an MT-oriented restriction to resolve syntactic ambiguity. The following is an example of rewriting a phrase:

> “Instruction and maintenance manuals” can be re-written to “instruction manuals and maintenance manuals”.

- *Do not omit the subject and the object of a sentence.*

This restriction is quite important for translation from Japanese, in which subjects and objects are very commonly omitted, to European languages. Today’s MT systems are not able to recover the omitted subjects and objects reliably.

- *Eliminate redundant words and expressions.*

Redundant expressions specific to a language, which often do not have counterparts in other languages, tend to produce awkward translations. Two Japanese examples are given below (Underlined parts are redundant).

> この (this) マニュアルは (manual *wa)* システム管理者を (system administrator *wo)* 対象とした (intended for) もの (thing) である (be)
> 最適化を (optimization *wo)* はかる (plan)

- *Avoid passive voice when the agent is explicit.*
- *Put a subordinate clause expressing a condition before the main clause.*
- *Avoid splitting an infinitive by an adverb unless the emphasis is on the adverb.*

Although these three restrictions may improve the readability for humans, violating them does not necessarily cause difficulties in MT systems.

3.3 Restrictions on style

Some restrictions on style have a good effect on machine translation. For example, use of a bulleted list eliminates a complex coordinate structure, and results in shorter sentences or phrases. Standardized usage of punctuation will also reduce the ambiguities in a sentence.

4 Controlled-language authoring support

Writing texts in a controlled language imposes a big burden on authors. Conscious attention to restrictions or writing rules interrupts their thinking. Moreover, they often cannot judge if their texts conform to the controlled language. Even if they notice that an expression violates a restriction, they may not find an alternative expression. Accordingly, controlled language authoring tools are strongly required.

Research and development of controlled language authoring support has been intensively conducted over the past decade. A representative project was the SECC (A Simplified English Grammar and Style Checker/Corrector) Project, funded by the European

Commission, in which the SECC checker/corrector was developed on the basis of the METAL machine-translation system [12]. LANT Ltd. has commercialized a controlled-language checker LANT MASTER based on the experience of the SECC Project. Other commercial controlled-language checkers include MAXit of Smart Communications, Inc. and ClearCheck of Logica Carnegie Group Inc. These commercial checkers, which are based on AECMA SE or their own controlled languages, provide a facility for user customization. A number of companies have also been developing authoring tools for in-house use. For example, Boeing developed Boeing Simplified English Checker (BSEC) [13] and is extending it to Enhanced Grammar, Style and Content Checker (EGSC) [14].

A controlled-language checker is a program that detects violation of restrictions and outputs alarm messages. The checker can be evaluated according to recall and precision ratios. Difficulty of detection varies greatly depending on the type of restrictions. Limitation on the sentence length can easily be checked. It is also easy to check the vocabulary if a word is either approved or unapproved independently of the meaning. However, it is difficult to check the vocabulary if a word is either approved or unapproved depending on the meaning. Some kinds of restricted expressions and structures can be detected reliably by shallow analysis of the text and pattern matching. An experiment with the SECC checker, which attained 93% recall and 87% precision, was reported [12]. On the other hand, detection of syntactic ambiguities is problematic. It is unavoidable that too many ambiguities are detected by a checker lacking semantic knowledge [15].

5 Future directions of controlled languages

Current problems in controlled languages are:

- to clarify the domains, other than procedural documents, to which controlled languages are effectively applied,
- to establish a method and tools for designing a controlled language for each domain, and
- to enhance the functions and precision of controlled-language authoring tools.

A key to solving these problems will be the corpus-based natural-language-processing techniques. Namely, the corpus-based paradigm has made marked progress in the 90s, and presently a lot of electronic texts are available in each potential domain in which a controlled language is applied. Analysis of the text corpus of a domain will clarify the applicability of a controlled language to the domain. Beyond being used for studying the vocabulary, a corpus will also play essential roles in designing an acceptable and effective controlled language. For example, corpus-based word-sense disambiguation will help us specify approved and unapproved meanings of polysemous words. Moreover, the capability of controlled-language authoring tools for detecting ambiguities can be greatly improved by using knowledge extracted from the corpora of domains. Another promising approach to authoring support is a tool that stores examples of rewriting and retrieves similar examples. This is analogous to the translation memory.

Finally, we should mention text annotation, which is regarded as an alternative to controlled languages or as an extension of controlled languages. CMU's KANT machine translation system, for example, uses a set of SGML tags to simplify source text analysis [16]. Proposals for text annotation such as TEI (Text Encoding Initiative) contain a versatile set of linguistic tags, including parts of speech, lexical meanings, phrase attachment, coordination, and anaphoric reference. These tags can substitute for some of the restrictions constituting a controlled language. So the role and scope of a controlled language should be reconsidered in light of that of text annotation.

References

[1] C. K. Ogden. Basic English, A general introduction with rules and grammar. Paul Treber & Co. Ltd., London (1932).

[2] Dictionary for Caterpillar Fundamental English. Caterpillar, Inc. (1974).

[3] P. J. Pym. Pre-editing and the use of simplified writing for MT: An engineer's experience of operating an MT system. Translating and the Computer 10, pp. 80-95, ASLIB (1990).

[4] G. Adriaens and D. Schreurs. From CORGAM to ALCOGRAM: Toward a controlled English grammar checker. Proc. of COLING-92, pp. 595-601.

[5] A guide for the preparation of aircraft maintenance documentation in the international aerospace maintenance language – Issue 1. AECMA document PSC-85-16598(1995).

[6] C. Kamprath, E. Adolphson, T. Mitamura, and E. Nyberg. Controlled language for multilingual document production: Experience with Caterpillar Technical English, Proc. of CLAW-98, pp. 51-61.

[7] L. Means and K. Godden. The Controlled Automotive Service Language (CASL) Project, Proc. of CLAW-96, pp. 106-114.

[8] I. Almqvist and A. Sagvall Hein. Defining Scania Swedish – A controlled language for truck maintenance, Proc. of CLAW-96, pp. 159-165.

[9] P. Goyvaerts. Controlled English, curse or blessing? – A users perspective, Proc. of CLAW-96, pp. 137-142.

[10] A. Bernth. EasyEnglish: Preprocessing for MT, Proc. of CLAW-98, pp. 30-41.

[11] A. Lehtola, J. Tenni, and C. Bounsaythip. Definition of a controlled language based on augmented lexical entries, Proc. of CLAW-98, pp. 16-29.

[12] G. Adriaens and L. Macken. Technological evaluation of a controlled language application: Precision, recall, and convergence tests for SECC, Proc. of TMI-95, pp. 123-141.

[13] R. H. Wojcik, P. Harrison and J. Bremer. Using bracketed parsers to evaluate a grammar checking application, Proc. of ACL-93, pp. 38-45.

[14] R. H. Wojcik and H. Holmback. Getting a controlled language off the ground at Boeing, Proc. of CLAW-96, pp. 22-31.

[15] H. Kaji. Language control for effective utilization of HICATS/JE, Proc. of MT Summit II, pp. 72-77 (1989).

[16] T. Mitamura and E. H. Nyberg, 3rd. Controlled English for knowledge-based MT: Experience with the KANT System. Proc. of TMI-95, pp. 158-172.

Note: CLAW-96 and -98 stand for the 1st and 2nd International Workshops on Controlled Language Applications that were held in Leuven in March 1996 and in Pittsburgh in May 1998, respectively.