Annotating Documents by Their Intended Meaning to Make Them Self Explaining: An Essential Progress for the Semantic Web

Hervé Blanchon and Christian Boitet

Laboratoire CLIPS BP 53
38041 Grenoble Cedex 9, France
{herve.blanchon, Christian.boitet}@imag.fr

Abstract. A Self-Explaining Document (SED) is a document enriched with annotations keeping track of all possible interpretations with respect to a given grammar and dictionary, as well as disambiguating choices. If disambiguation is complete and has been done by the author himself, a SED conveys "the author's intention". The availability of SEDs might considerably reduce misunderstanding between authors and readers, and perhaps lead to the assignment of a "meaning certification level" to any part of a document. We present ways to integrate these annotations into an arbitrary XML document (SED-XML), and to make them visible and usable to readers for accessing the "true content" of a document. We also show that, under several constraints, a SED, once translated into a target language L, might be transformed into an SED in L with no human interaction. Hence, the SED structure might be used in multilingual as well as in monolingual contexts, without addition of human work.

1 Introduction

We first proposed the concept of Self-Explaining Document in [5] as an answer to some question raised while experimenting with a new incarnation of the interactive translation paradigm, the DBMT approach (Dialogue-Based Machine Translation) in the LIDIA project [6]. We observed (again) that translation introduces ambiguities which are not present in the source text. Traduttore, traditore… For example, the two French words "remplacer" (replace by a new thing) and "replacer" (put back into place) were both translated by "replace" in English. It also happened that all disambiguated analyses of a sentence produce the same translation, as ambiguous as the original. One example was the translation from French into Russian of the famous sentence «the man sees the girl in the park with a telescope».

This raised an objection to DBMT: what is the use of disambiguating the source text if ambiguities reappear in the translation(s), or even worse if new ones are created? Would it not be better to try and produce translations which preserve the ambiguities, and dispense with Interactive Disambiguation (ID) altogether?

Unfortunately, the experience of human translation shows that ambiguities can be exactly preserved only in some cases, and that to do it purposefully is quite difficult and often leads to unnatural ways of expression in the translation. It is also quite clear

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 601 – 612, 2006.

that the "transferable" ambiguities vary with the target language. Finally, although some texts may be intentionally ambiguous, especially in poetry and politics, we take it that the vast majority of ambiguities are not intentional, but are due to the intrinsic nature of natural languages. Of course, some authors write more clearly than others, but all authors write unambiguously in any programming language, unambiguous by construction, and ambiguously in any natural language, ambiguous by nature!

This motivated the idea of Self-Explaining Documents: if the source and target documents are accompanied by their (unambiguous) linguistic structure, with the indications of potentially ambiguous parts, and if the reader in the target language may obtain a clarification of unclear parts in a user-friendly way, the objection disappears. As human users are notably not very sensitive to ambiguities, however, we should find a way to warn the reader that the target text is ambiguous.

2 DBMT: A Context for SED Production

After having worked in the direction of "suboptimization" for 15 years [10], we turned to high quality Dialogue-Based Machine Translation. DBMT is a new paradigm derived from that of Interactive MT (IMT) and geared to various translation situations where other approaches, such as the Linguistic-Based (LBMT) and the Knowledge-Based (KBMT) approaches, are not adequate.

In DBMT, although the linguistic knowledge sources are still crucial, and extralinguistic knowledge might be used if available, emphasis is on indirect pre-editing through negotiation and clarification dialogues with the author in order to get high quality translations without revision.

Authors are distinguished from "spontaneous" writers or speakers by the fact that they want to produce a "clean" final message and may be willing to enter into such dialogues. The crucial difference with usual IMT is that interactive disambiguation is not performed during an analysis or transfer process, but on a "multiple" data structure factorizing all possible analysis results. Hence, the author is not "slave of the system", but decides if and when s/he wants to enter the disambiguation dialogue.

The first situation considered (in 1990) was the production of multilingual technical documentation in the form of HyperCard[1] documents. A page of such a document is a card. A card may contain different kinds of objects such as graphics, buttons and textual fields. The linguistic MT lingware was based on multilevel transfer with interlingual acceptions, properties and relations implemented in ARIANE-G5.

The first mockup, LIDIA-1 [6], demonstrates the idea on a HyperCard stack presenting short ambiguous French sentences in several disambiguating contexts. This document is translated into three documents, German, Russian and English. Although this mockup does not implement all features of the general design — a complete implementation would have called for considerably more human resources than were available — it demonstrates the potential of the approach.

[1] HyperCard is a Macintosh-based (MacOS-7 to MacOS-9) environment for the production of hypertextual documents called "stacks".