Abstract Machine Translation of Interactive Texts

Mary Flanagan Manager,
Applied Research
CompuServe
mflanagan@csi.compuserve.com

Online texts can be grouped in three general categories. **Reference** texts include news and magazine articles, software technical support texts and general reference materials such as encyclopedia excerpts. Reference texts are written for publication to a large and varied audience and are generally static or infrequently changed. Most reference texts are well written and grammatical and contain few spelling or punctuation errors. They cover a range of topic areas and contain a diverse but stable vocabulary. **Communicative** texts, such as e-mail and bulletin board messages consist typically of casual discussions. Most communicative texts are directed at an individual or a small audience. Communicative texts often contain sentence fragments, misspellings and misused punctuation. Topics are varied but the largest vocabulary category is online jargon rather than any set of subject-specific terms. Communicative texts are more transient than reference texts. They are typically available for a short period of time or are written for a single reading. They are often copied and modified with comments. **Interactive** texts are online chat or other real-time online communications. Interactive texts may be directed at an individual in a chat room or the members of the room in general. A high percentage of interactive text is sentence fragments. Topics are fluid, often changing every few lines as participants enter and leave the chat room. Interactive text is highly elliptical and contains many misspellings and punctuation anomalies. Interactive texts also contain many personal names. Online jargon is overwhelmingly the largest vocabulary category and includes many abbreviations and chat-room specific terms. Interactive texts are exchanged in real time or near real time.

Interactive texts pose unique challenges for machine translation. Because most MT systems assume the sentence as the unit of analysis, interactive texts, with their low percentage of complete sentences, present parsing difficulties for MT systems and often yield low quality output. Topic shifts, spelling and grammar errors and ellipses compound the difficulty. Automated pre-editing and dictionary development can improve output quality. From the perspective of translation production, there are also challenges in delivering real-time, high volume translations. User interface is one such challenge. Translations delivered directly into a chat room are confusing to participants, and increase the rate at which text scrolls off. An alternative is to maintain separate chat rooms for each language, however this segregates participants by language and tends to reduce cross-language interaction. Timely delivery and maintenance of conversations threads is also difficult. Typical chat room activity consists of multiple overlapping conversations. The translation must maintain the discussion thread in the same order as in the source, and translations must be delivered within seconds to preserve the coherence of the original.