Submitted by: University of Maryland Center for Advanced Study of Language
Presenter: Erica Michael
Additional Contributors/Authors: Sergey Blok, Michael Bloodgood, Petra Bradley, Ryan Corbett, Michael Maxwell, Peter Osthus, and Paul Rodrigues, Benjamin Strauss
Topic: Evaluating Parallel Corpora: Assessing Utility for Use with Translation Memory Systems in Government Settings

Translation memory (TM) software allows a user to leverage previously translated material in the form of parallel corpora to improve the quality, efficiency, and consistency of future translation work. Within the intelligence community (IC), one of the major bottlenecks in implementing TM systems is developing a relevant parallel corpus. In particular, the IC needs to explore methods of deploying open source corpora for use with TM systems in a classified setting. To address this issue we are devising automated metrics for comparing various corpora in order to predict their usefulness to serve as vaults for particular translation needs. The proposed methodology will guide the use of these corpora, as well as the selection and optimization of novel corpora.

One of the critical factors in TM vault creation is optimizing the trade-off between vault size and domain-specificity. Although a larger corpus may be more likely to contain material that matches words or phrases in the material to be translated, there is a danger that some of the proposed matches may include translations that are inappropriate for a given context. If the material in the vault and the material to be translated cover similar domains, the matches provided by the vault may be more likely to occur in the appropriate context. To explore this trade-off we are developing and implementing computational similarity metrics (e.g., n-gram overlap, TF-IDF) for comparison of corpora covering 12 different domains. We are also examining summary statistics produced by TM systems to test the degree to which material from each domain serves as a useful vault for translating material from each of the other domains, as well as the degree to which vault size improves the number and quality of proposed matches.

The results of this research will help translation managers and other users assess the utility of a given parallel corpus for their particular translation needs, and may ultimately lead to improved tagging within TM systems to help translators identify the most relevant matches. Use of open source materials allows tool developers and users to leverage existing corpora, thus holding the promise of driving down costs of vault creation and selection. Optimizing vaults also promises to improve the quality, efficiency, and consistency of translation processes and products.