(U) For Media Mining, the Future Is Now! (conclusion)

FROM: Joseph Picone and
Human Language Technology (S23)
Run Date: 08/07/2006

(S//SI) Media Mining Across a Wide Range of Languages

(S//SI) One of the challenges in deploying this Media Mining HLT is the need to cover the very broad range of languages. Unfortunately, most of the languages of interest to the Agency are not of interest to commercial concerns because they are not likely to be profitable, and businesses run on profit.

(S//SI) Though COTS products such as NEXminer have covered commonly-taught, "dense" languages such as English and Spanish, and have made great inroads lately into a few less-commonly-taught languages and dialects found in the Middle East, it is unclear that any COTS product will ever cover the vast inventory of languages that NSA analysts are required to understand. Therefore, the HLT PMO is developing an enhancement of this Media Mining technology that can process over 90 languages using a combination of language-specific and universal phones. This agency capability, developed within R64, the Human Language Technology Research Group, is known as Universal Phonetic Recognition (UPR).

(S//SI) New languages can be easily added to the technology by drawing on Agency linguistic knowledge of a language combined with publicly available language resources. As world events shape our language needs, UPR provides a way to respond within minutes to new language needs, for example to support the GWOT.

(U) IVE: Technology that Can Separate the Wheat from the Chaff

(S//SI) A second, equally important enhancement under development is the ability for this HLT capability to predict what intercepted data might be of interest to analysts based on the analysts' past behavior. Much like the way in which popular sites like amazon.com are able to track and predict buyer preferences, integration of Intelligence Value Estimation (IVE) on both SRI and message content, offers the promise of presenting analysts with highly enriched sorting of their traffic. Imagine if you came to work each day knowing that the best five intercepts needing transcription were sitting at the top of your queue waiting for you.

(S//SI) Of course, such Media Mining IVE capabilities need not be limited to SRI and key word searches. In collaboration with S202B, Analytic Technologies for the Enterprise, the HLT PMO Media Mining team is also developing new metadata analysis capabilities based on language, speaker, gender, and dialect identification, presenting this information to analysts through conventional query tools such as UIS. Advanced programs like RT-10 are integrating other forms of information, such as geospatial coordinates. RT-10 will also send automatic alerts to analysts when incoming intercept meets certain search criteria.

SERIES: (U) HLT

(S//SI) Voice $_{RT}$ will soon be integrated with standard Agency voice tools such as UIS and HOTZONE. Analysts will be able to configure the tool via the web, and access scores on their traffic using NUCLEON.

(U) Bringing it All Together

(S//SI) The integration of these technologies into an automated system will bring two major innovations: faster response time and improved productivity. Our challenge goal is to "index, tag, and graph" all incoming intercept, and this will soon be within reach. Using HLT services, a single analyst will be able to sort through millions of cuts per day and focus on only the small percentage that is relevant. The amount of collection can be increased orders of magnitude without further stressing the analyst population, allowing the Agency to cast a much wider SIGINT net and taking in a much richer catch.

(S//SI) And again, the power of HLT is truly realized through integration of multiple SIGINT technologies. In the future, we will further develop technologies such as word search to support cross-lingual queries. Sites that lack expertise in a given language will be able to issue queries in English and receive results translated from the target language back into English. This marriage of word search and Machine Translation has great potential as a force multiplier. Mapping meaning and tradecraft across languages will be a key challenge here.

(S//SI) Similarly, because a search term will be tagged with a "semantic class identifier," such as "place name," it will be relatively straightforward to integrate this technology with the Enterprise Knowledge System (EKS) and allow sophisticated capabilities such as social network analysis to operate on voice content. In the HLT PMO long-term vision, analysts will be able to construct complex queries, such as, "Where is the mayor of Baghdad?" or "Show me all the intercept containing information about explosive devices that occurred yesterday in the downtown area of Baghdad near the Al-Rashid Hotel," and obtain answers directly in English, or in their foreign language if they prefer, with a link to the documents containing the answers.

(U//FOUO) **We are entering a golden age for HLT.** Powerful and inexpensive computers, high-speed networking, and advanced algorithms are being combined to revolutionize the analyst desktop.

(U//FOUO) For more information about these capabilities, please contact the HLT PMO office ("go HLT" or call (s) [redacted] .
