Ai Translate
by LSI

Optimized MT for Multi-Language Discovery and Investigation

By Jason E. Boro, Esq., Linguistic Systems, Inc.

Foreign-language document translation for eDiscovery and Investigation can be costly and time-consuming. This use-case validates the significant benefits of 'Optimized MT' as a core element of the Electronic Discovery Reference Model.[1]

Companies have a duty to conduct their business as though they may be sued. They must preserve data that eventually may be relevant as evidence at a trial. And they should be prepared to consider legal action that may be necessary to protect shareholder interests or to preserve the value of business assets.

The legal requirement to preserve potentially relevant electronically stored information (ESI) is a cost of doing business.[2] However, judges are likely to admonish parties if they allow discovery costs to get out of control.[3] The cost of carrying out a suit or a defense is supposed to be proportionate to the scope of the potential damages.[4]

CURRENT MODEL: HIGHER COST, HIGHER RISK

When litigation becomes necessary, it is customary to assemble a team to review collected data in order to furnish an early case assessment. If foreign language documents are involved — thousands, at times — the conventional practice has been to staff the team with bilingual contract attorneys. But this model can be impractical and risky for the following three reasons:

1. It may be very time-consuming to staff a bilingual attorney team.
2. After staffing delays, haste may result in data-entry or data-recovery errors.

3. The use of foreign-language attorneys often results in a cost per reviewed document that is three times higher than using English-speaking attorneys.[5]

THE CASE FOR OPTIMIZED MT

An alternative to using bilingual contract attorneys is the use of Optimized Machine Translation (OMT). This process can achieve "dramatic savings."[6] It was applied, for example, to a review of a large number of Japanese-language documents. The law firm that used the English output of optimized machine-translated Japanese documents (*instead of Japanese-language source documents read by bilingual attorneys* to conduct the review) concluded that the optimized-machine translations were "of sufficient quality to allow them to be coded by U.S.-based contract reviewers with no Japanese language skills."[7] The cost savings was estimated to be one-half compared to using bilingual attorneys to read the source Japanese documents.[8]

Linguistic Systems deploys Optimized MT to help clients significantly reduce their eDiscovery costs and turnaround-time. It involves three action steps:

1. Developing glossaries of (potentially) thousands of case-specific custodian names, product names, acronyms, and keywords that improve the MT output;
2. Running several individual language-specific MT passes against the foreign-language documents using the improved engine (see Step 1). These would start with the dominant language engine, then the engine for the second most prevalent language, and so forth until all languages are converted to English;
3. Deploying a team of bilingual subject-matter domain specialists to post-edit any highly responsive machine-translated materials.

Selecting attorneys for their legal knowledge, rather than for their foreign-language ability, allows legal providers to assemble *more qualified* case teams. Attorneys will be able to review Optimized MT output more quickly than foreign-language reviewers who are not as familiar with the case.

Leading eDiscovery vendors who handled large foreign-language matters using MT have shown that English-speaking case teams can be more efficient. Working with Optimized MT, they gleaned legal insights that achieved case objectives at a lower cost versus a team of less-experienced reviewers native to the source languages who reviewed non-converted materials[9].

EXAMPLE OF USING OPTIMIZED MT FOR eDISCOVERY

Here is how Linguistic Systems implemented one Optimized MT solution. A multinational technology client faced a massive investigation in Russia involving 159 custodians. The workflow proceeded as follows:

- The client and the eDiscovery vendor collected 8 terabytes (TB) of electronic documents along with 1.3 TB of user-generated data. The resulting dataset comprised 10,080,503 documents in English, Russian, and other foreign languages.
- To remove non-relevant documents, deduplication techniques were used in the first culling phase to remove 2,800,918 documents.
- Next, the eDiscovery vendor used term-by-term reporting to evaluate and refine the client's search terms — thereby excluding another 5,118,472 documents.
- Applying junk-file analysis, data collection was reduced by another 1,477,855 documents.

- In this manner, the original 10 million documents were reduced by 93%, culling the data down to 683,258 relevant documents.

Ai TRANSLATE PROCESS

Ai Translate by LSI™ (formerly Select Translation Service)[10] is a best-in-class language translation solution launched nearly a decade ago by Linguistic Systems. In one tool, it offers access to 7,500 skilled certified translators, 64 statistical machine translation (SMT) engines, 16 neural machine translation (NMT) engines, ISO-certified security, and up to 6 translation quality options. Users can access Ai Translate by LSI™ as a managed service, as a plugin within the Relativity eDiscovery platform, or via the Web. — Source: AiTranslate.com.

Linguistic Systems used **Ai Translate** as follows:

- Relevant documents were machine-translated first for the dominant language, Russian. The next most-frequently identified language was translated second, and so forth. This approach converted all relevant documents from their source language(s) into English.
- Linguistic Systems then worked with the client to develop a robust glossary of non-machine-translatable words. They included certain conceptual terms, repeated phrases, proper nouns, personal names, place names, and industry acronyms. Using several iterations in an A.I. mode, this step boosted the accuracy and comprehensibility of the machine-translated output.
- As a result, the client could respond more efficiently to the opposing party's discovery requests. The client ultimately was able to *exclude 99.5%* (all but 314 files) from its potentially relevant cache of 74,492 Russian documents.
- Linguistic Systems then performed a light post-edit (LPE) on the 314 files — those on which the review remained focused. This treatment option saved

the client 60% (compared to traditional human translation), while facilitating compliance with eDiscovery. The difference in cost between converting the documents to English using MT with a light post-edit versus traditional human translation amounted to a savings of $83,000.[11]

Ai TRANSLATE BY LSI™ CONSULTING

Translating thousands, even hundreds of thousands, of foreign-language documents to find the most vital information is a daunting challenge. The path to success can vary from project to project.

Ai Translate Consulting offers a deeper opportunity to collaborate with experts in systems, processes, and languages to achieve an optimal result. Linguistic Systems has developed the breadth and quality of its' capabilities by serving the AmLaw 100, Fortune 100, and tens of thousands of other clients over the course of more than 50 years.

Each of the following steps can enhance the performance and ROI of a substantial eDiscovery foreign-language translation project:

- **Customization of the machine translation engine** (MTE) by ingesting aligned relevant examples of foreign- and English-language sentences. Directed machine-learning trains the engine to produce more accurate, more consistent output.
- **Ownership of the customized MTE.** Once a client invests in MTE customization, that asset is available for subsequent similar cases in the same industry and designated language. An initial investment in customization of the MTE yields ongoing benefits.
- **Categorization and prioritization of massive document streams.** Source language identification, relevance categorization, conversion prioritization, and

translation-method selection are all critical tasks that **Ai Translate by LSI™ Consulting** teams help facilitate.

- **Asset allocation**. Applying the most suitable translation and review assets appropriately can attain the most cost- and time-saving results.
- **Targeted utilization of resources**. Clients selectively flag certain documents for higher-quality translation treatments. After receiving post-edited files, case teams may assign foreign-language attorneys to contribute their review capability to the workflow.
- **Custom security implementations. Ai Translate** can handle millions of documents, securely, from one application. Authorized client and legal representatives and translators work in a secure, cloud-based environment. Even so, a custom security solution may be required, and **Ai Translate by LSI™ Consulting** teams can work with clients to customize translation-system security.

OPTIMIZED MT AS THE FIRST CHOICE

This use case shows how a collaborative team — consisting of client, legal counsel, and eDiscovery — and language-services consultants — can leverage an Optimized MT solution to improve the quality, timeliness, and ROI of a major eDiscovery effort. As evidence:

- Clients are able to surgically isolate a small subset of relevant data that may require full human translation and review. This often includes certification of translated materials needed for admission in court.
- The approach reduces the amount of time from data collection up to the review-phase of the EDRM. It reduced the culling time — for example, by enabling a client (such as Paul Hastings, in another published case study)

— to quickly "identify relevant materials for further review and to seek hand translation of particular documents as necessary."[12]

- Machine learning and glossary assets become portable. A client may re-use them on subsequent matters.
- **Ai Translate's** language identification functionality assists in normalizing data to a single language (English). This simplifies custodial control over the materials. Meanwhile, having the ability to comprehend the content allows the law firm to control decision-making regarding the case.
- Files are translated in the cloud within a secure loop that encloses the language services provider's and the client's data on the MT servers.
- Relieved of the demands of translation management, case-team members can re-focus on their core responsibilities of legal work.

CONCLUSION

Optimized MT may be a highly cost-effective and time saving alternative to bilingual review for situations with a high volume of documents, originating in one or more foreign languages, in connection with discovery. By isolating a much smaller subset of MT-converted documents that require additional translation scrutiny, case teams can *significantly reduce* translation costs while reducing the risks associated with a lawyer-linguist approach.

APPENDIX: COST CALCULATIONS

This use case involved a multi-national technology company that faced a large investigation in Russia. For that client, Linguistic Systems machine-translated 74,492 Russian documents. Using 4,000 to 5,000 documents per GB[13] as a divisor, we calculate that the subset of Russian documents may have totaled approximately 14.9 to 18.6 GB of the totality of 8 TB of custodian data. The cost of the MT for this volume would be approximately (14.9 X 2000 X 0.75) to (18.6 X 2000 X 0.75) ➔ or, approximately $22,350 to $27,900. *See* **Table 1.**

Glossary enhancement allowed the client to reduce the file-count of significant relevant documents to just 314 files. Then, a light post-edit consisting of 553,241 words was performed on the machine-translated documents, costing 553,241 X ($0.10 / word) = $55,324.

The cost for attorneys to review 314 culled files — which were MT-converted to English after enhancing the MT with glossary and light post-edit — would have been approximately 314 X ($0.55 per doc) to 314 X ($0.64 per doc) = from $172 to $200. *See* **Table 2.**

Meanwhile, a traditional bilingual review would present certain challenges outside of the **Ai Translate** platform. First, there is the challenge of identifying the comprised languages before reviewing the 74,492 Russian documents.[14] The minimum cost for the attorney labor assumed in this scenario would be $2 per document reviewed, or about $148,984. *See* **Table 2.**

The Optimized MT workflow required translation and eDiscovery phases: MT conversion (at a cost in the range of approximately, $22,350 – $27,900); light post-editing (costing about $55,324); and focused English-language attorney review (costing about $172 to $200). Thus, the Linguistic Systems solution (with

rounding) cost between $77,847 and $83,425. Therefore, Optimized MT saved the client the difference between $148,984 and as much as $83,425, or approximately $65,000.

Table 1. # Docs. (of Different Types) per GB; Cost of MT Conversion Per GB

File-type	# Docs in 1 GB	Avg. KB Per Doc	# Atty. Review Hrs.	Cost of English-only Atty. + MT Conversion[15]
DOC	3541	282	**16**	$ 1,980 - $ 2,060
XLS	2530	395	**12**	$ 1,860 - $ 1,920
PPT	702	1425	**3**	$ 1,590 - $ 1,605
PDF	1893	528	**9**	$ 1,770 - $ 1,815
TXT	14291	70	**65**	$ 3,450 - $ 3,775
MSG	4873	205	**22**	$ 2,160 - $ 2,270
EML	11219	89	**51**	$ 3,030 - $ 3,285

Table 2. Bilingual vs. English-Speaking Attorney Reviewer Cost ($ per Doc.)[16]

Language Groupings	FR, DE, IT, PT, ES	AR, BG, HR, CS, NL, IW, HU, LV, LT, NO, PL, RO, RU, SK, SL, SV, TR	ZH, JA, KO	DA, EL, ET, FI, MT
English-only Atty. ($ / doc reviewed)	$ 0.55 - $ 0.64			
English-only eDiscovery contract attorney pay-rate (@ $30 - $35 / hr.)[17]				
Bilingual Atty. ($ / doc reviewed)	$ 1.45 - $ 2.36	$ 2.00 - $ 2.55	$ 1.82 - $ 3.64	$ 2.73 - $ 2.91
Bilingual eDiscovery contract attorney pay-rate ($) / hr.[18]	$ 40 - $ 65	$ 55 - $70	$ 50 - $ 100	$ 75 - $ 80

END NOTES

[1] *See* Margaret Rouse's post, recovered at URL: https://searchcompliance.techtarget.com/definition/EDRM-electronic-discovery-reference-model (*stating* the definition, "The Electronic Discovery Reference Model (EDRM) is a framework that outlines standards for the recovery and discovery [. . .] of digital data.")
[2] *See* Federal Rules of Civil Procedure ("FRCP") Rule 37(e) (*defining* the duty to preserve ESI, *i.e.*, in anticipation of litigation).
[3] *See e.g.*, The State of E-Discovery 2018: A Survey of Industry Trends, Practices, and Challenges Faced, at Part Three: Challenges Faced, *q.v.*, Slide 35 of 43 -- recovered at URL: https://www.exterro.com/state-of-eDiscovery-2018/ (*highlighting*, therein, relevant trends, *inter alia*, "Proportionality gaining traction"; "Legal right v. practical ability issues are being resolved"; "And judges themselves are adopting a more active role in achieving 'just, speedy, and inexpensive resolutions'.").
[4] *See* Federal Rules of Civil Procedure ("FRCP") Rule 26 (*standardizing* proportionality requirements in civil discovery).
[5] For example, bilingual attorneys cost more than English-speaking attorneys – by an average factor of 3 times – for French, German, Italian, Portuguese, or Spanish; 3.8 times – for Arabic, Bulgarian, Croatian, Czech, Dutch, Hebrew, Hungarian, Latvian, Lithuanian, Norwegian, Polish, Romanian, Russian, Slovak, Swedish, or Turkish; 4.5 times – for Chinese, Japanese, or Korean; and 4.8 times – for Danish, Greek, Estonian, Finnish, or Maltese. *See* **Table 2.**
[6] *See* Catalyst Case Study, "Catalyst Optimized machine translation Can Reduce the Cost of Japanese Document Review by 50%," (*observing*, "Whereas the cost to use bilingual, Japanese-fluent reviewers would be a minimum of $630,000, the cost using OMT and U.S. reviewers would be half that, $315,000.")
at 2nd para., at 2nd sent. – recovered at URL: https://catalystsecure.com/pdfs/case_studies/Catalyst_CaseStudy_Catalyst_Enhanced_Machine_Translation_Can_Reduce_the_Cost_of_Japanese_Document_Review_by_50_Percent.pdf.
[7] *See* Id.
[8] *See* Id.
[9] *See e.g.*, "Was Samsung Deal a Watershed for Use of Machine Translation in FTC Second Requests?" by Bob Ambrogi, May 15, 2012, posted in Catalyst Language Services, E-Discovery Search Blog, at 2nd art., recovered at URL: https://catalystsecure.com/blog/category/catalyst-language-services/.
[10] *N.b.,* AiTranslate by LSI was formerly branded as Select Translation Services (STS).
[11] The use case involved light post-edit of 553,241 English words, at a cost of $55,324.10 *versus* traditional human translation treatment, *i.e.*, which would have been at a cost of $138,310.30.
[12] *See Op. Cit.*, "Was Samsung Deal a Watershed for Use of Machine Translation in FTC Second Requests?" by Bob Ambrogi, May 15, 2012, posted in Catalyst Language Services, E-Discovery Search Blog, at 2nd art., at "Using MT for the Second Request," at para. 2, at sent. 2, recovered at URL: https://catalystsecure.com/blog/category/catalyst-language-services/ (*arguing* that Optimized MT was as effective as more expensive solutions, given that the objectives of conducting eDiscovery in the first instance were all served by means of the machine-translated materials, *i.e.*, notwithstanding their not having "the same level of quality as the hand-translated materials.")
[13] *See* "How Many Documents in a Gigabyte? Revisiting an E-Discovery Mystery," by John Tredennick, August 20, 2015, at Conclusion, last sent., at p. 6 of 6, recovered at URL: https://catalystsecure.com/blog/2015/08/how-many-documents-in-a-gigabyte-revisiting-an-e-discovery-mystery/.
[14] *N.b.*, the corpus of Russian documents represented a subset of the reviewable population that prior steps of the culling process – performed in consultation with vendors – enabled the client to isolate from the universe of potentially-relevant documents.

[15] The **Table 1** calculation of the cost of MT assumes a unit charge of 0.75 per 500 KB file, as a putative industry average.

[16] Based on New York City rates related by a trusted contact, we assume an average review rate for Bilingual eDiscovery attorneys of 27.5 docs / hr., and we assume an average review rate of 55 docs / hr. for English-only eDiscovery attorneys. *C.f.,* Chris Egan & Glen Homer, "Achieve Savings by Predicting and Controlling Total Discovery Cost," Metropolitan Corp. Couns. (Dec. 1, 2008), http://www.metrocorpcounsel.com/ pdf/2008/ December/ 08.pdf, *cited in* David Degnan, "Accounting for the Costs of Electronic Discovery," 12 Minn. J. J. Sci. & Tech. 151 (2011), at fn.74. Available at: https://scholarship.law.umn.edu/mjlst/vol12/iss1/7.

[17] The English-only eDiscovery attorney pay-rate of $30 - $35 / hour assumed herein is based upon information conveyed by a trusted industry contact, as confirmed by posted rates.

[18] For the purposes of illustrating the comparative cost-basis of carrying out two different foreign-language review strategies, with the use case we make a broad assumption that there are four price levels of language-expertise for bilingual review of original materials, on the one hand, or light post-editing treatment of machine-translated materials, on the other. *E.g.*, our Eastern U.S. partners report observed bilingual contract attorney pay rates -- for these groupings – ranging as follows: *Group 1* ($40 - $60 / hr.); *Group 2* ($55 - $70 / hr.); *Group 3* ($50 - $100 / hr.); *Group 4* ($75 - $80 / hr.). We assume that Russian language falls into Group 2 of this illustrative stratification.