A Technology Agnostic Approach to Machine Translation: Machine Translation Summit XIV

Lori Thicke
LexWorks Ltd.
Lori.Thicke@lexworks.com

LexWorks is the machine translations services branch of Lexcelera, widely known as one of the leading experts in technology-agnostic machine translation. Nearly a decade's experience in full production environments has shown that machine translation is not one-size-fits-all. LexWorks carries out deployments, engine training and post-editing using all of the main approaches: SMT (online and server based), RBMT and Hybrid.

This engine-agnostic approach to machine translation is critical because no single approach – SMT, RBMT or Hybrid – suits all content types, all projects, and all languages. Best-of-breed solutions can be identified by applying a set of practical guidelines based on factors such as language combination, content type, file formats and available data as well as by rigorously benchmarking engine performance at project launch.

LexWorks has developed a systematic process to identify the highest performing engine in a given use case scenario by using the criteria below for guidance. These criteria form 'rule of thumb' assumptions to test with rigorous benchmarking of quality scores obtained when comparing each of the various engine-types on the content in question.

Content Type & Other Considerations	Online SMT	Hybrid	RBMT	SMT
Documentation, reports, online help, UI		✓	✓	
FAQs, forums, UGC	✓			✓
Patents, other broad domain	✓			✓
Marketing materials				
Insufficient in-domain/out-of-domain data	✓	✓	✓	
Poor grammar, spelling	✓			✓

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 309–311.

Language Considerations (Sample)	Online SMT	Hybrid	RBMT	SMT
French, Spanish, Italian	✓	✓	✓	✓
Russian, Japanese, German		✓	✓	
Norwegian, Danish, Thai	✓			✓

As another decision aid in determining the best-of-breed engine, comparing features of Rules-Based versus Statistical performance in several areas, yields the following:

Area	Feature	RBMT	SMT
Capability	Number of languages handled out of the box	~20	~50
Capability	Add rare language pairs		✓
Cost	Free or Open Source version exists	✓	✓
Cost	SaaS models exist	✓	✓
Quality	Output is fluent		✓
Quality	Can handle bad grammar		✓
Quality	Significant quality improvements with pre-editing	✓	
Quality	Output is predictable	✓	
Quality	Uses specified terminology applying correct grammar	✓	
Quality	Handles software tags without special programming	✓	
Quality	Can be integrated with TM tools	✓	✓
Suitability	Better performance with UGC and broad-domain content (e.g. patents)		✓
Suitability	Better performance for on-the-fly translations of short shelf-life content		✓
Suitability	Better performance for documentation/UI	✓	
Suitability	Suited to rare language pairs		✓

Suitability	Suited to full post-editing with improvements made to engine in near real-time	✓	
Training	Learns automatically		✓
Training	Rapid improvement cycle	✓	
Training	Effective with limited training corpus	✓	

In the LexWorks process, assumption testing based on benchmarking uses a variety of quality metrics such as BLEU, GTM, SymEval, plus human sentiment analysis, understandability measures and, in the case of online customer support uses, answers to the question: “Did this solve your problem”?

On the enterprise side, other technology agnostic users of MT include Adobe, Autodesk, PayPal and Symantec, all of whom, like LexWorks, believe that a good MT strategy looks for the best-of-breed solution on a case-by-case basis.