Xtensyon.

Multilingual Enterprise Search: Handling Mixed-Language Queries with Consistent Relevance

August 26, 2025 • Xtensyon Labs • 7 min read

Search quality drops when users mix languages, abbreviations, and technical terms in one query. This brief covers indexing, normalization, and evaluation methods that improve recall without punishing precise keyword searches.

TL;DR

- Collect real mixed-language queries as a test set before tuning anything.
- Normalize carefully and keep original text for highlighting and exact matches.
- Evaluate by intent match and downstream actions, not just click-through.
- Maintain a synonym list owned by teams; it matters more than model swaps.

Executive Summary

Many enterprise teams search the way they speak and write. That means connectors in one language, technical terms in another, and local abbreviations blended together. Traditional keyword search struggles, while embeddings alone can over-generalize. We outline a practical approach that improves recall while keeping relevance stable: language-aware normalization, hybrid retrieval, and a small evaluation set that reflects real usage. The goal is finding the right internal document fast and with confidence.

Why It Matters

When search fails, people ask in chat, forward old PDFs, or keep private copies. That creates version drift and increases operational risk. Mixed-language search improves productivity and reduces reliance on informal channels, especially for operations, support, and policy-heavy teams.

What We Built

- A query normalization step that handles common connectors and particles without stripping meaning.
- Hybrid retrieval that combines keyword matching with embeddings for better coverage.
- A curated list of synonyms, acronyms, and domain terms maintained by power users.
- An evaluation suite of mixed-language queries with expected source documents.

Observed Outcomes

- Higher recall for mixed-language queries without reducing precision for keyword-heavy searches.
- Less repeated searching once relevance improved for common abbreviations and internal jargon.
- Faster onboarding for new team members because internal documents were easier to find.

Implementation Notes

- Do not over-normalize. Keep original terms for exact matches and highlighting.
- Track "no result" and "reformulated query" rates as leading indicators.
- Start the synonym list small and review it monthly with domain owners.
- Use a reranker only after the basics are stable and measurable.

Governance & Risk

- Do not store raw queries with identifiers without a retention plan.
- Make synonym changes auditable. Treat it as production configuration.
- Avoid indexing personal drives without clear ownership and access rules.

Release Checklist

- Do we have a real-world mixed-language test set?

- Is retrieval hybrid with clear fallbacks?

- Are synonyms owned and reviewed regularly?

- Do we track no-result and query reformulation rates?

- Are logs stored with privacy controls?

Conclusion

Multilingual search is less about chasing the newest model and more about listening to how teams ask questions. Start with real queries, tune normalization, and measure outcomes. The improvements show up quickly.

Keywords

multilingual search   mixed-language queries   embeddings   information retrieval   enterprise search

evaluation