IBM Watson
Language Translator

Best Practices Guide

Questions:
Zia Mohammad – z.mohammad@ibm.com
Watson Language Translator: Tools & Community

**LT Homepage:** [https://www.ibm.com/watson/services/language-translator/](https://www.ibm.com/watson/services/language-translator/)
See the Developer Tools section of the Language Translator webpage

**LT on Stack Overflow:** [https://stackoverflow.com/questions/tagged/ibm-watson-cognitive](https://stackoverflow.com/questions/tagged/ibm-watson-cognitive)
Read questions and get answers from other Watson developers

**Code Libraries & SDKs:** [https://github.com/watson-developer-cloud](https://github.com/watson-developer-cloud)
NodeJS | Python | Swift | Java | Unity | .Net | Salesforce

**API Reference:** [https://www.ibm.com/watson/developercloud/languageTranslator/api](https://www.ibm.com/watson/developercloud/languageTranslator/api)
Learn how to effectively call the Language Translator API

**Getting Started:** [https://console.bluemix.net/docs/services/language-translator/getting-started.html#language-translator](https://console.bluemix.net/docs/services/language-translator/getting-started.html#language-translator)
See the tutorial and the demo | Learn how to use your own data | See the release notes & interact with the developer community
Best Practice

Translating Between Languages (L1 to L2)

Tips & Tricks:

1. Make sure that your selected language and inputted language are the same. Does the language you selected match the text?

2. Be sure to de-tag your input text while removing any mark-ups.

3. If LT's sentence segmentation is missing some sentence breaks in your input, use double newlines in the input text to force sentence breaks.

4. Use more-basic versions of characters over more-exotic variants in Unicode.

For example:
Using the few most-standard quotation-mark characters -- the basic non-directional " or the most-basic directional ones open-quotes and close-quotes, “ ” and « », is more advisable than more-obscurc versions.

Similarly for single quotes and apostrophes, regular hyphens vs. "non-breaking", regular spaces vs. special spaces like non-breaking, wide, and narrow, etc.
Tips & Tricks Continued:

5. Unicode often has multiple representations of characters which look alike.

   For example:
   The "full-width" version of the Latin alphabet, variants on commas and semicolons, etc.
   Having problems? Try a more-basic version of the character, an English version, or one which is specifically compatible with L1.

6. Favor composed characters over combining versions.

   For example:
   A single Unicode character which represents a letter with an accent mark is more likely to yield good results than the character which represents the unaccented letter and a "combining" accent-mark character.

7. Beware of weakly-directional characters, like numerals intended to be taken left-to-right but appearing among right-to-left text such as Arabic.
Model Customizations:

1. Have domain specific documents that might exemplify vocabulary, phrasing, and text patterns? Try using a forced glossary or corpus customization for domain specific translations.

   **For example:**
   Think of smoked fish. If LT translates "smokehouse" into a phrase which means "house of smoke", and the target language has a better word for "smokehouse",.

   If this is the core of what a customer's business does, pair "smokehouse" with the target language's word for "smokehouse" in a forced glossary, so as to ensure that top-level fundamental documentation of what the company does remains well-written.

2. Parallel-corpus customization picks up the variations in language and punctuation specific to an organization or industry. The more a style of writing differs from news stories and other generic factual information sources, the more benefit parallel corpus customization is likely to provide compared to the LT service's base model for a language pair.
Model Customizations Continued:

3. What are some examples of domain-specific material which would improve customization:

**Words:** Specialized meanings (Brand names, CEO names, technical jargon, hyphenated pair vs. two words).

**Punctuation:** Customer-support chat etc. may not end in punctuation or have capitalized words, even in languages like German where nouns are capitalized and identically spelled verbs are not.

**Grammar:** News and factual sources are focused on formal, polite statements of fact and occasionally questions. If any of the following occur regularly in domain, they should be included in customization:

*Informal language* such as tu vs vous (French), du vs Sie (German)

*Instructions (Imperatives):* "do this/Machen Sie" "go/Gehen Sie" vs "You are doing" "You are going" (different word order and/or inflections)

*Types of language* that would take an exclamation point at the end

*Questions* if any are typical of the domain and unlike the questions that would appear in news
Training Models and Separating Data:

1. Separate your available in-domain data into training data (90%) and a development test set (10%). Avoid duplicating copies of the same sentence in both the training and development test set. Select the development test set from the full range of content you wish to support. If you selected the last 10% of a data file, it might all be from one subdomain or source.

2. IBM invented BLEU, the industry standard automatic metric for MT performance: [https://en.wikipedia.org/wiki/BLEU](https://en.wikipedia.org/wiki/BLEU)


**Note:** For character-based languages like Japanese and Chinese, run BLEU at character level by adding spaces between the characters.

**Run:** Send your development test set through translation twice:

**a.** Use the standard base model for your language pairs.

**b.** Use your customized model. Compute a BLEU score for each translation. Successfully customized models yield a higher BLEU score than a base model. Common gains range .5 BLEU point - few BLEU points.
Watson Language Translator: Continue the conversation

**New Releases:** https://console.bluemix.net/docs/services/language-translator/release-notes.html#release-notes
See the what’s new with Watson Language Translator | Have you tried out Neural Machine Translation?

**IBM developerWorks Blog:** https://developer.ibm.com/dwblog/
Head over to the Cognitive Computing section to learn more

**IBM Watson YouTube Channel:** https://www.youtube.com/channel/UCxPJljXUHvUd9idyfEHvXqg
Watch inspirational videos and see how developers are using Watson

**Slack Community:** wdc-community.slack.com
Join our channel and take part in the ongoing conversation