Using Bigrams and Unigrams as Translation Units Seems to Work Better

**Motivation**
- No natural word boundaries in Chinese and Japanese: need to determine the index unit first.
  - Using word segmentation
  - Cutting sentence into n-grams
- Both types of indexes have been used in translation units in CLIR
- Question: How do they compare as translation units in CLIR?

**Using different index units**
- W (Word): Sentences are segmented into words.
- U (Unigram): Sentences are cut into single characters.
- B (Bigram): Sentences are cut into overlapping bigrams of characters.
- WU (Word and Unigram): Sentences are segmented into both words and single characters.
- BU (Bigram and Unigram): Sentences are cut into both overlapping character bigrams and single characters.
- B+U: Interpolating Bigram(B) and Unigram(U)

**Problems of word segmentation in information retrieval**
- Segmentation Ambiguity: 
  - 发展中（developing）/国家(country)
  - 发展（development）/中(middle) 国家(country)
  - 发展（development）/中国(China)/家(family)
- Different words may have the same or related meaning, especially when they share common characters.

**LM approach for monolingual IR and CLIR**
- KL-divergence between query language model and document language model

**Using different translation units**
- Translates English Words to Chinese Words(W), Unigrams(U), Bigrams(B), or Bigrams&Unigrams(BU).

**Conclusion**
- N-grams produce results comparable to the average results of NTCIR6 in Chinese, Japanese, and Korean.
- For Chinese
  - N-grams are generally as effective as words for monolingual IR.
  - For Cross-language IR, N-grams approaches can be even better than dictionary-based word translation
- N-grams can be interesting alternative indexing and translation units to words.
- Worth further investigations.