I. Finding candidate anchors and target links

We extracted all anchor links in the Wikipedia collections. Each anchor link is composed of the surface text (mention) and its target link. Mentions may be different from the title of the target topic page and a mention is often shared among different concepts. For an input article, we do mentions matching first to mind a bag of possible target pages. In order to mind the links to concepts that didn’t exist in CJK Wikipedia but in English Wikipedia, we utilized POS tagging technique. The surface texts that match POS pattern for anchor texts will be translated into English. More possible linked pages can be found by matching the translated text to all titles in the English Wikipedia collection.

II. Computing relevance between cross-lingual Wikipedia pages

There are two approaches to compute similarity score between CJK Wikipedia pages and English Wikipedia pages. *Keyword Similarity*: We use the mentions in mention table of English Wikipedia as word list to calculate the similarity. We translated the input Chinese or Japanese article to English by machine translation system first. Then apply mention matching to both the translated article and target linking page. The score is given by Dice’s coefficient.

$$\text{keywordSim} = \frac{2 \times \text{A \& B}}{|\text{A}| + |\text{B}|}$$

A. keywords of input article
B. anchor texts of candidate Wikipedia page

*LDA model*: LDA is a model introduced by Blei et al (2003), designed to automatically induce latent hidden topics from discrete data. Each LDA topic is a distribution over the words of the corpus. Documents are represented as a mixture of topics. This is to say that every topic of the model has a probability in every document, and that the similarity between two documents can then be calculated as a similarity between the topics composing it. A new translated English document is first converted in its bag of words vector, and then to the distribution over the LDA topics.

The comparison between two documents is done with a cosine similarity between the two documents topic vectors.

$$\cos(A, B) = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2 \times \sum B_i^2}}$$

A is a vector representing a document of the English Wikipedia, B is the vector representing the original Chinese or Japanese input document after its translation into English.