

# NTHU at NTCIR-10 CrossLink-2: An Approach toward Semantic Features

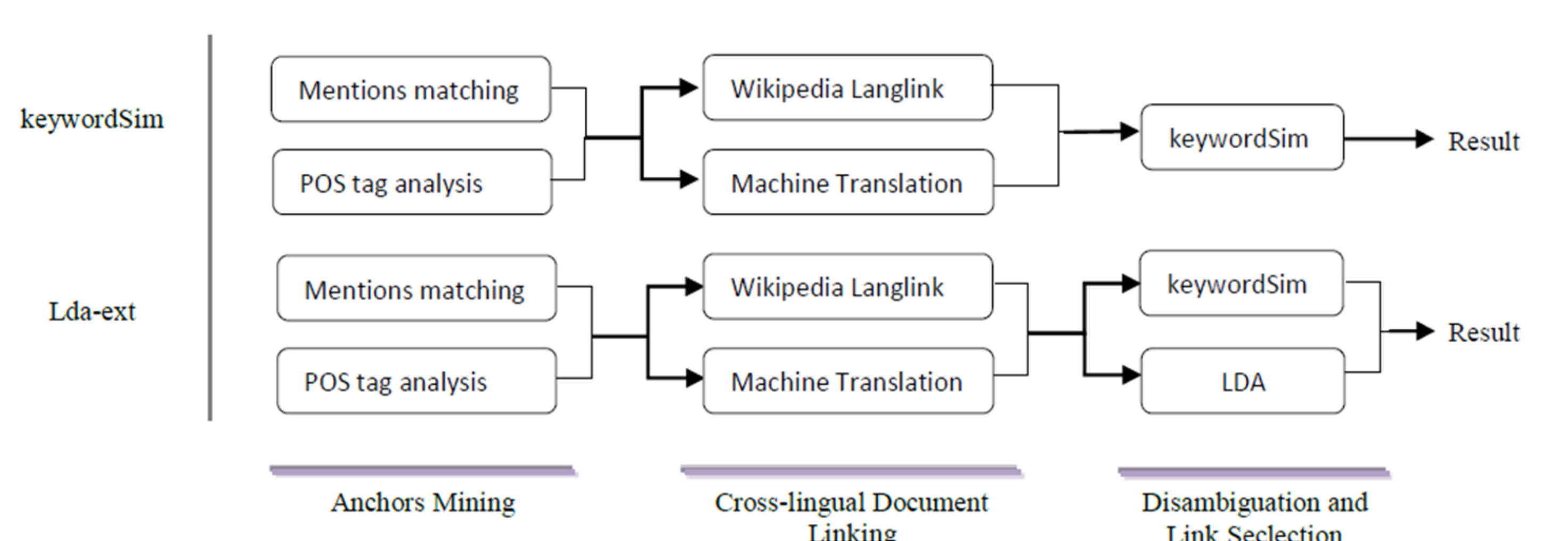


Yu-Lan Liu  
Department of Computer Science  
National Tsing Hua University,  
Taiwan  
ikulan12@gmail.com,

Joanne Boisson  
Institute of Information Systems  
and Applications,  
National Tsing Hua University,  
Taiwan  
joanne.boisson@gmail.com

Jason S. Chang  
Department of Computer Science,  
National Tsing Hua University,  
Taiwan  
jschang@cs.nthu.edu.tw

## System Overview

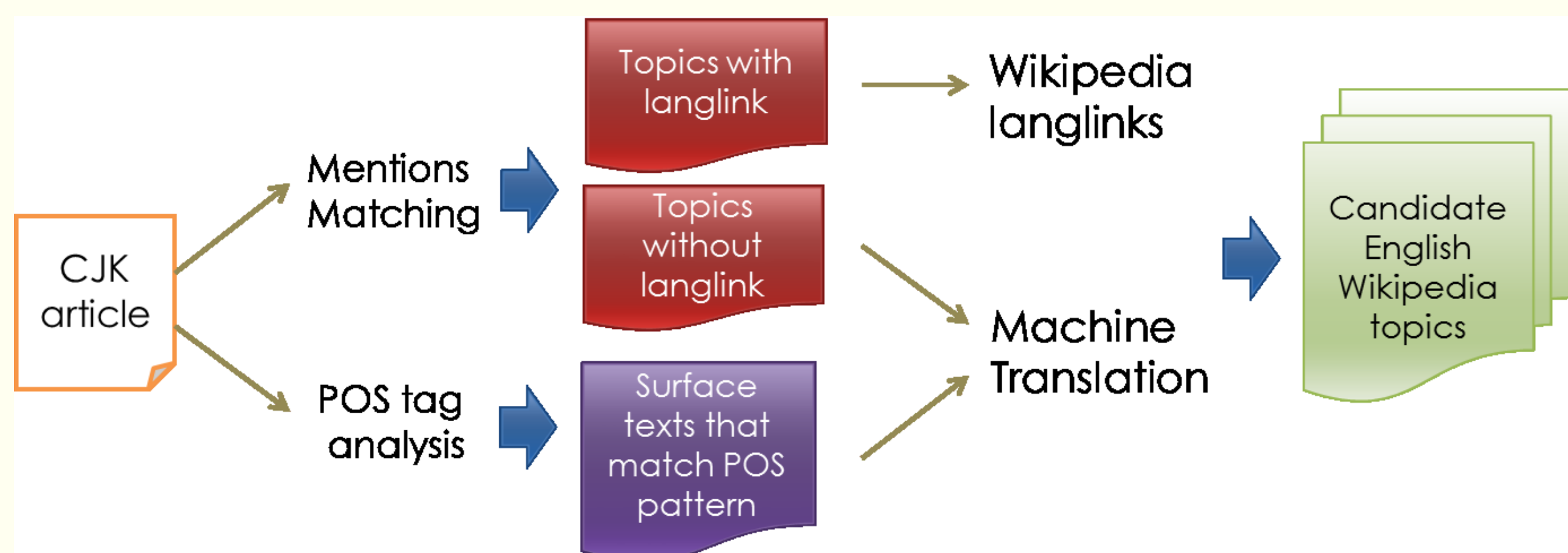


## Method

### 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 bag 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.

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

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.

$$\cos(A, B) = \frac{\sum_{i=1}^n A_i \cdot B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \cdot \sqrt{\sum_{i=1}^n (B_i)^2}}$$

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.

racing car engine race cars driver motor formula engines speed  
 la el mexico spanish puerto san del juan mexican chile  
 government patrolling court accused police act law clerk defending security  
 regiment army polish infantry battalion brigade division poland battle  
 album song chart band track vocals albums songs guitar single  
 navy ship naval ships hms royal officer vessel uss admiral  
 river lake antarctic island km park glacier mountain dam mountains  
 orchestra piano opera composer symphony czech violin dgg jazz  
 al ottoman khan armenian muhammad pakistan muslim empire afghanistan Israeli  
 church bishop catholic cathedral rev diocese ordained college parish priest

The ten most probable words obtained for ten LDA topics

## III. Anchors and links selection

In this stage, we rank all the anchor text and target link pair candidates discovered previously by the combination of many measures. The features are listed below:

**Global Keyness of anchor text (gk):** The score presents how likely the n-gram being an anchored.

**Category Probability of target page (cat\_p):** Some categories are very likely to be anchored, e.g. Countries, Movie players. We analyzed all anchor links in Wikipedia collection and computed the portion of categories.

**Parenthesis (pp):** Whether the anchor text is parenthesized or not.

**Keyword Similarity (keywordSim):** The relatedness of target page and origin page.

**LDA similarity (lda\_sim):** The relatedness of target page and origin page.

$$gk(t) = \frac{\text{show times for } t \text{ as anchor text}}{\text{total show times of } t}$$

$$CatP(d) = \sum_{c \in d} \frac{\text{the number of target documents that has category } c}{\text{total number of categories in all anchor links}}$$

## Experiment Result

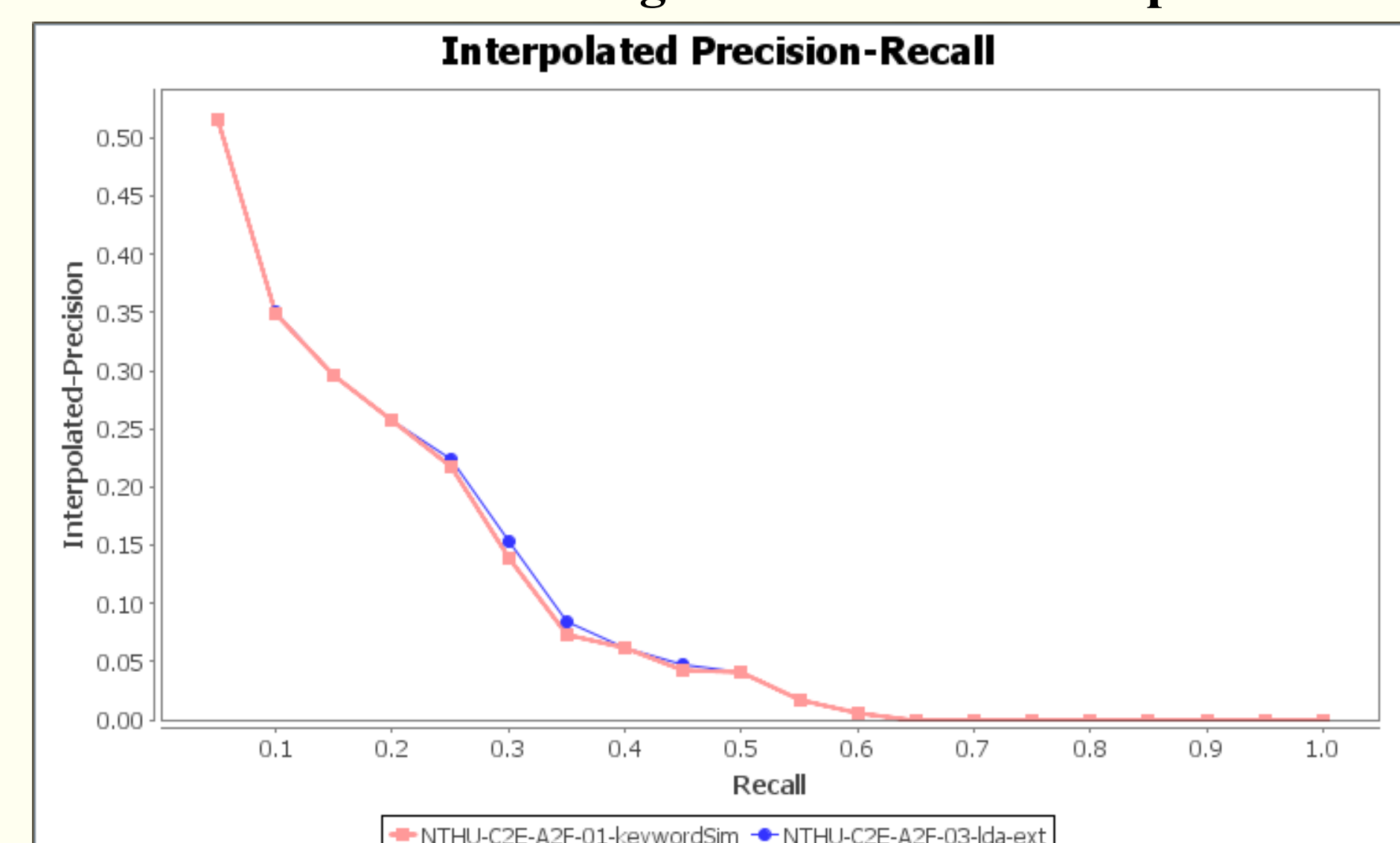
Table 4. Performance of NTHU's system in Chinese to English subtask

		LMAP	R-Prec	P5	P10	P30	P50	P250
F2F	Best Score	0.517	0.520	1.000	0.972	0.779	0.582	0.123
	Wikipedia							
Ground-truth	keywordSim	0.080	0.192	0.256	0.236	0.221	0.194	0.068
	Lda-ext	0.082	0.194	0.256	0.240	0.224	0.195	0.070
F2F	Best Score	0.069	0.180	0.384	0.368	0.320	0.266	0.123
	Manual							
Assessment	keywordSim	0.025	0.096	0.192	0.136	0.120	0.123	0.051
	Lda-ext	0.034	0.114	0.192	0.136	0.121	0.126	0.078
A2F	Best Score	0.113	0.147	0.096	0.072	0.083	0.082	0.064
	Manual							
Assessment	keywordSim	0.012	0.024	0.040	0.024	0.024	0.028	0.008
	Lda-ext	0.021	0.036	0.040	0.024	0.025	0.029	0.013

Table 5. Performance of NTHU's system in Japanese to English subtask

		LMAP	R-Prec	P5	P10	P30	P50	P250
F2F	Best Score	0.548	0.561	0.946	0.938	0.829	0.657	0.178
	Wikipedia							
Ground-truth	keywordSim	0.083	0.189	0.254	0.246	0.224	0.199	0.084
	F2F							
Manual	Best Score	0.312	0.418	0.520	0.460	0.357	0.267	0.066
	Assessment							
Manual	keywordSim	0.102	0.138	0.184	0.164	0.133	0.123	0.049
	Assessment							
A2F	Best Score	0.270	0.120	0.144	0.120	0.107	0.083	0.037
	Manual							
Assessment	keywordSim	0.127	0.074	0.064	0.068	0.064	0.062	0.017

### The F2F evaluation of Chinese to English result with Wikipedia Ground Truth



### The F2F evaluation of Japanese to English result with Wikipedia Ground Truth

