Wikification for Scriptio Continua

Yugo Murawaki Shinsuke Mori

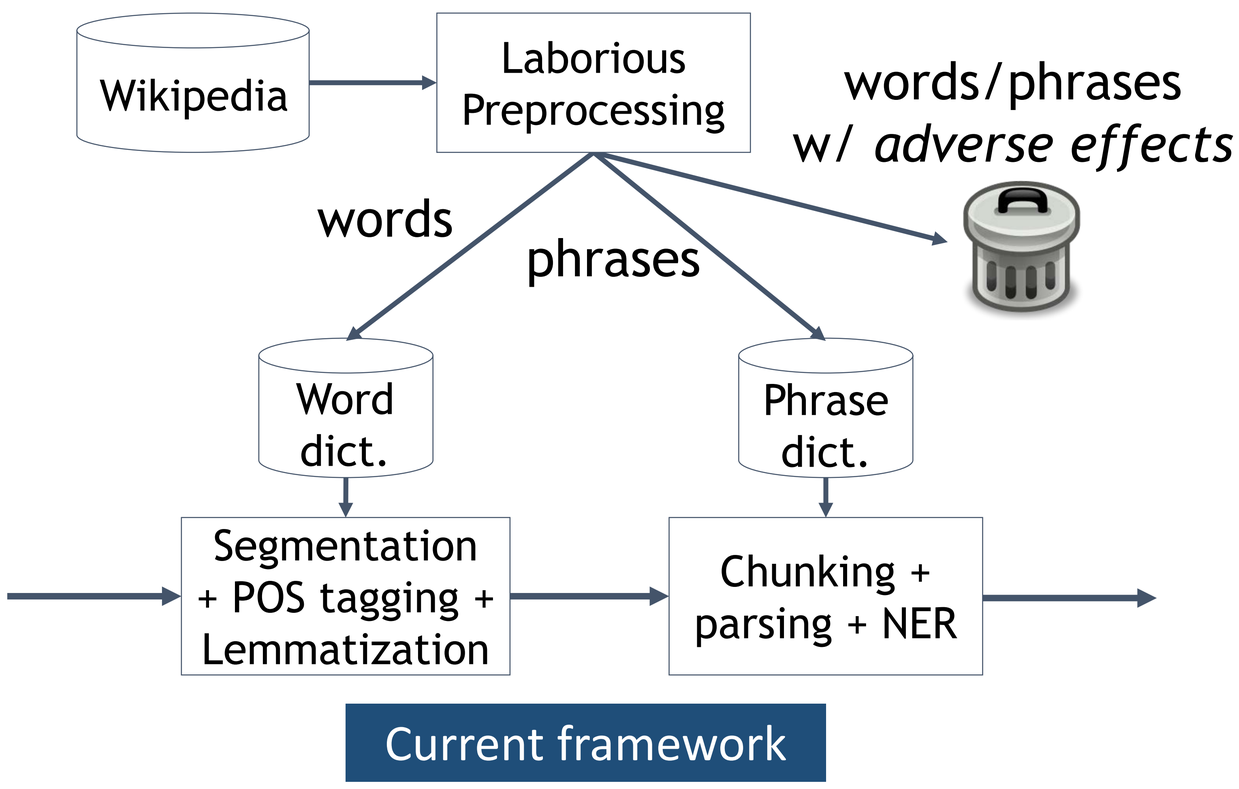
京都大学 KYOTO UNIVERSITY

Summary

- Designed the task of Japanese Wikification
- Constructed annotated corpora
- Implemented a simple wikifier
- Motivation: design an elegant framework of Japanese language analysis in the presence of heterogeneous lexical resources

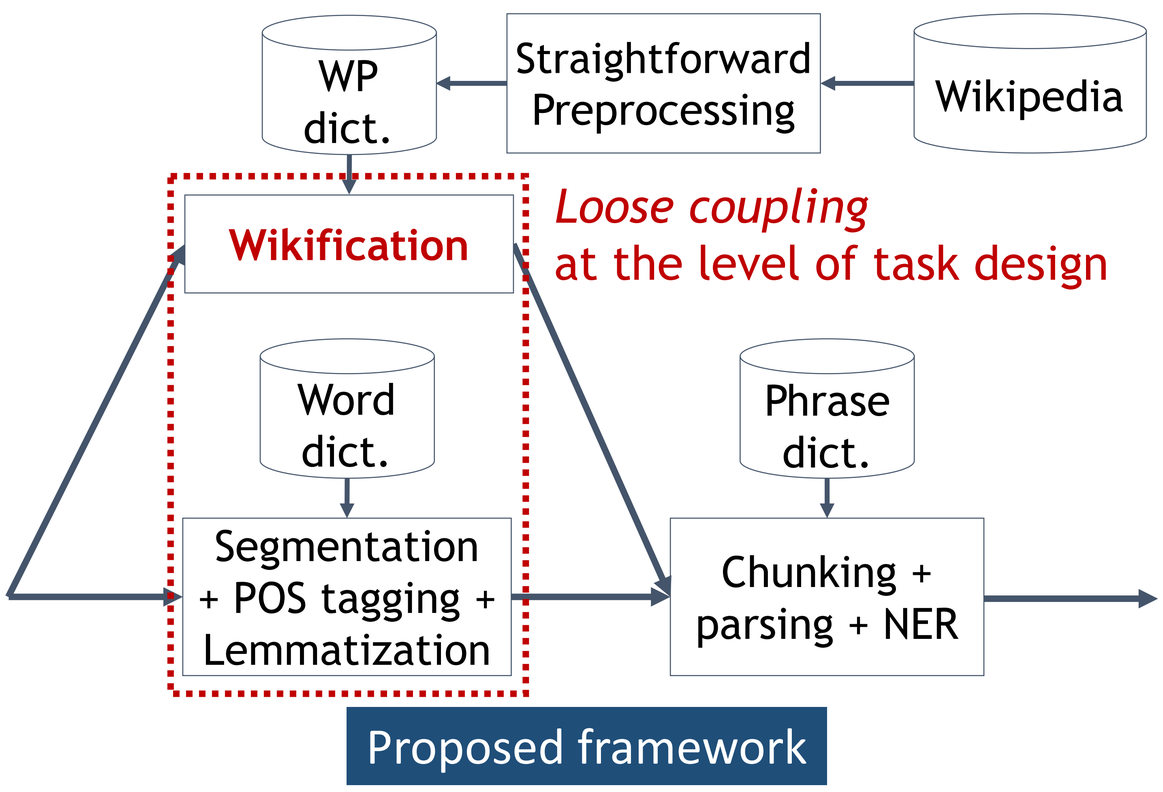
1. Motivation

- Japanese employs *scriptio continua* and requires word segmentation as the first step of NLP
- Lexical knowledge, encoded as a word dictionary, is crucial for high accuracy
- External lexical resources, like Wikipedia, are potentially useful but hard to be incorporated
  - Inherent segmentation mismatch
  - Need to avoid hardly predictable adverse effects in word segmentation



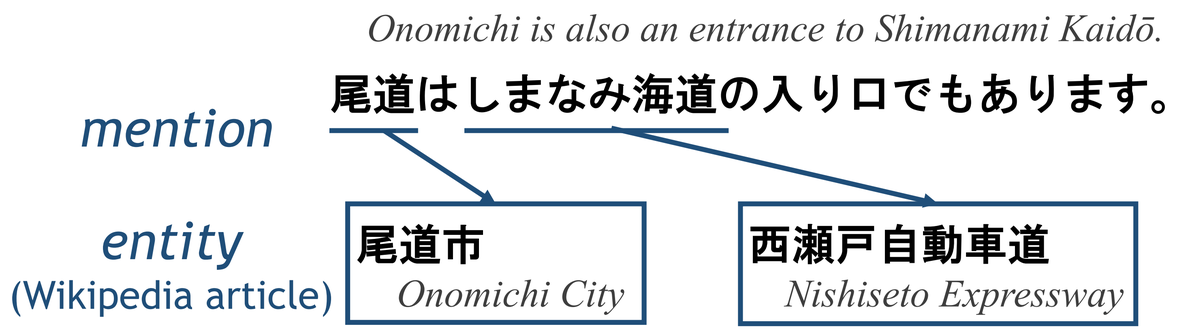
Solution:

- Design a separate task for an external resource
  - *Wikification* for the case of Wikipedia
- Loose coupling with word segmentation makes it easier to fully exploit Wikipedia



2. Wikification: Task Design

- Mention detection + entity disambiguation



To exploit Wikipedia as a lexical resource, we chose the following task specifications:

- No *NIL* detection for efficient candidate enumeration
- Cover any topics, not just named entities
- Exhaustive detection, not just important concepts
  - Pro: More consistent annotation
  - Con: Difficult to use Wikipedia itself for training
- Prefer longer mentions for consistency
- Allow topical matching that seems unavoidable
  e.g. mention *manufacturer* → entity *manufacturing*

3. Corpus Annotation

Three genres of texts

1. White papers (BCCWJ-OW)
2. Yahoo! Blog (BCCWJ-OY)
3. Twitter

Added another layer to multiple layers of annotation (segmentation, dependency, and POS)

gloss	*ID*	*PID*	*word*	*POS*	*wikification*
Onomichi	001	002	尾道	名詞	B 尾道市
TOP	002	009	は	助詞	
Shimanami	003	004	しまなみ	名詞	B 西瀬戸自動車道
Kaidō	004	005	海道	名詞	I
GEN	005	006	の	助詞	
entrance	006	007	入り口	名詞	
COP	007	008	で	助動詞	
too	008	009	も	助詞	
be	009	010	あ	動詞	
SUF	010	011	り	語尾	
POL	011	012	ま	助動詞	
SUF	012	013	す	語尾	
.	013	-1	。	補助記号	

Corpora	#Docs	#Sents	#Words	#Mentions	Mention detect.	Entity disambig.
BCCWJ-OW	8	504	23,592	3,486	.895	.861
BCCWJ-OY	34	509	9,239	922	.878	.832
Twitter	NA	2,942	37,009	3,649	.768	.739

- The last two columns show the performance of an independent human annotator (≒ upper bound)
  - Correlated with the formality of text
- A simple wikifier we created performed 5-12% below the human annotator
- Performance w.r.t. training data suggests data are enough at least for the simple wikifier