W82

Symmetric Correspondence Topic Models for Multilingual Text Analysis

Kosuke Fukumasu (*Kobe University*)
Koji Eguchi (*Kobe University*)
Eric P. Xing (*Carnegie Mellon*)

W82

Multi-type topic models

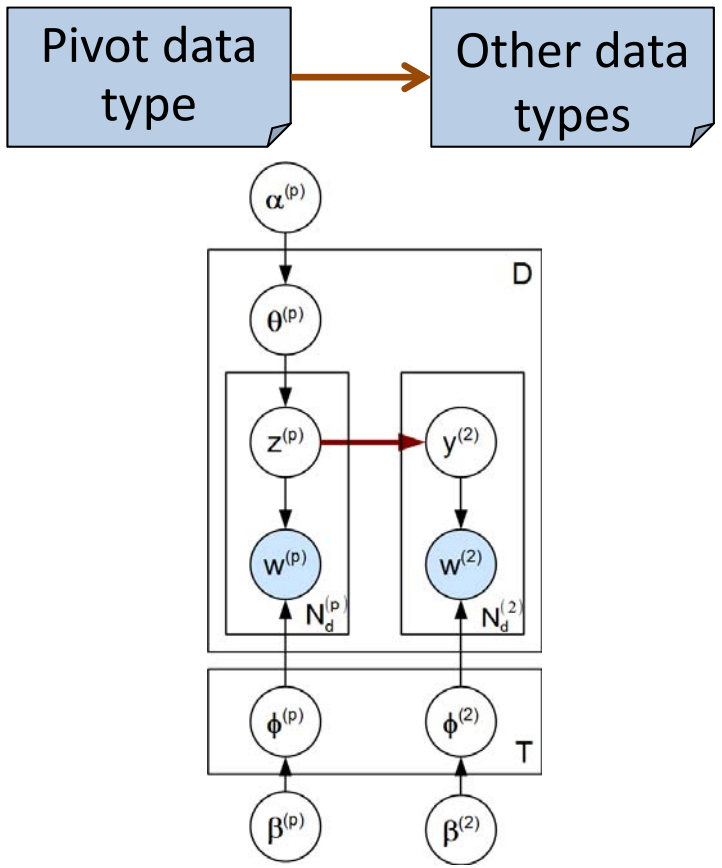
- Topic modeling approach has been applied to multi-type data, such as:
  - ✓ text-annotated image data
  - ✓ multilingual aligned text data
- Some prior models fail to capture dependencies between the data types.
- CorrLDA (Correspondence LDA) [Blei et al., 2003] does well with the dependencies; however, it requires an annoying constraint, which requires a ***pivot*** —which data type is generated first— to be specified in advance.

W82

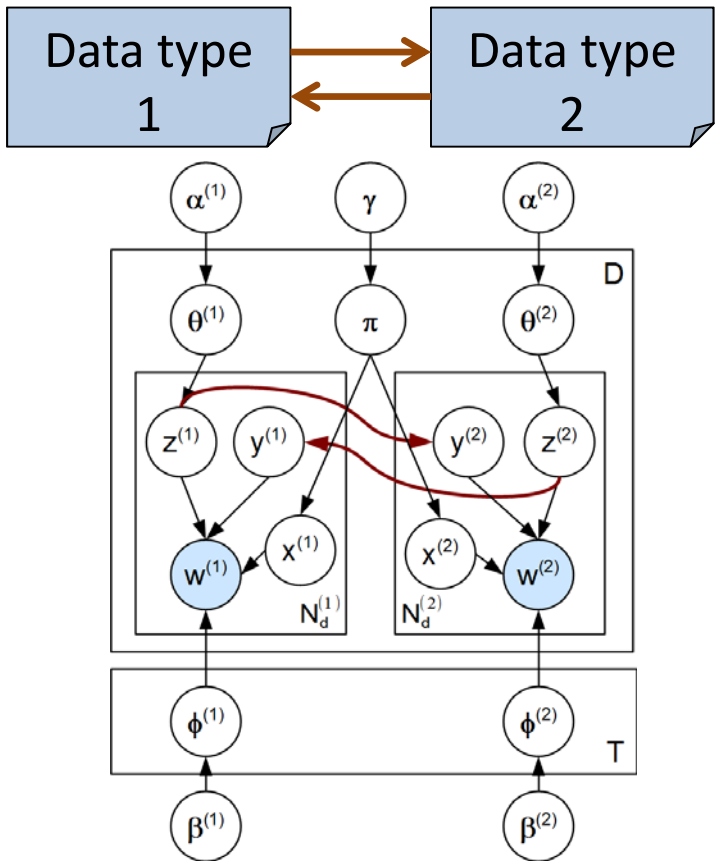
Symmetric Correspondence LDA

- SymCorrLDA can adjust the proportion of the pivot for each data type, using a switch variable.

CorrLDA

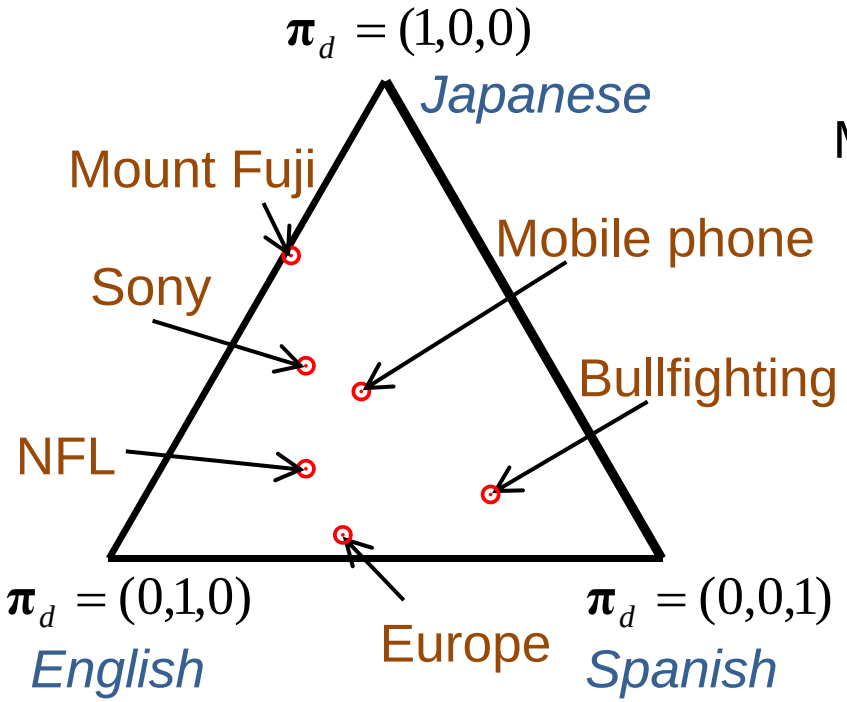


SymCorrLDA



W82

Results



Mean reciprocal rank in finding counter parts.

	Jap to Eng	Eng to Jap
CorrLDA	0.299	0.316
SymCorrLDA	**0.326**	**0.335**

9% improved. 6% improved.

(Both improvements are statistically significant.)

- We experimented with the multilingual aligned text data gathered from *Wikipedia*.
- CorrLDA works significantly more effectively than the other prior models.
- SymCorrLDA is more effective than CorrLDA.