Word Alignment with Synonym Regularization

Hiroyuki Shindo, Akinori Fujino, and Masaaki Nagata

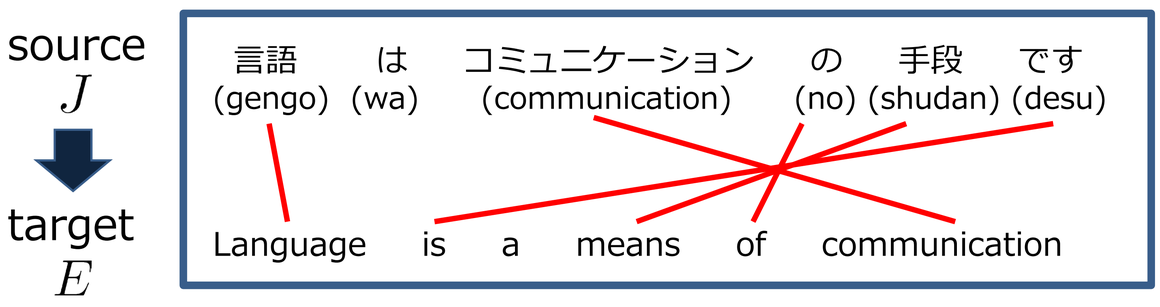
NTT Communication Science Laboratories

Overview

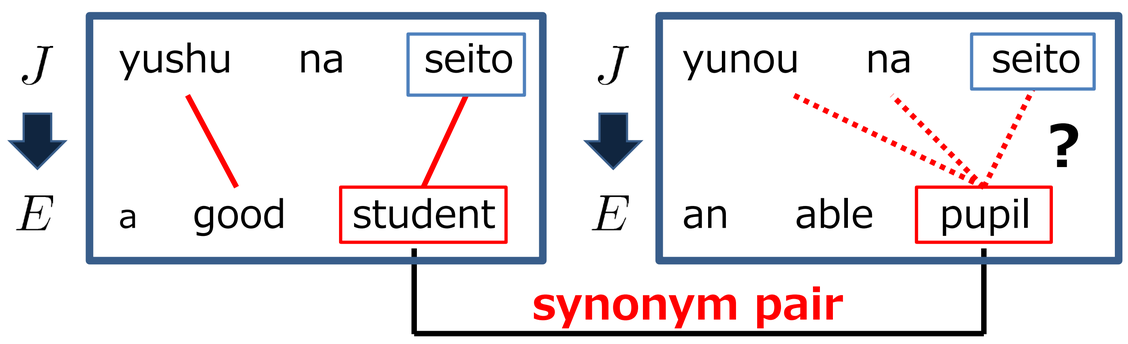
★ We propose a **Bayesian** **word alignment model** that incorporates **synonym knowledge** in bilingual corpus with **topic model** .

Background

Approach:
improve word alignment accuracy with synonym dictionary



Synonym information is helpful



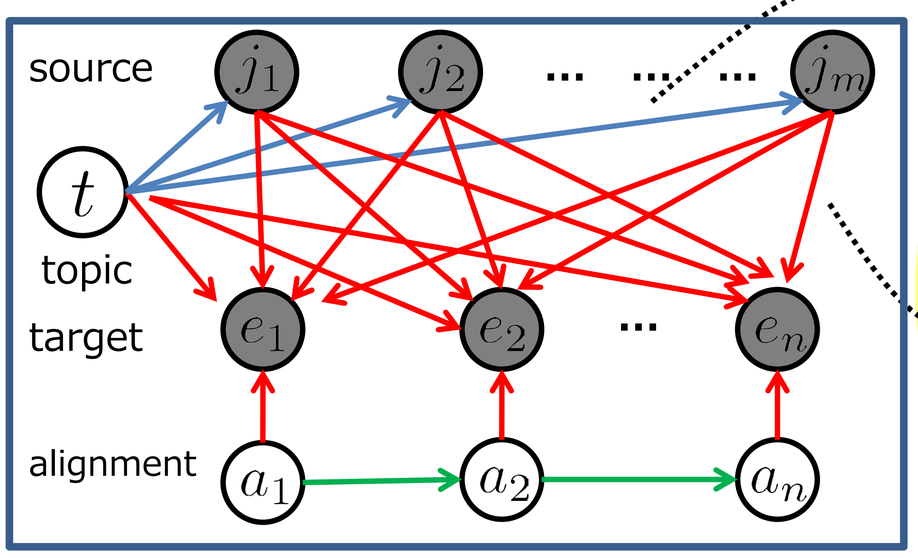
・When we know ‘student’ and ‘pupil’ are synonym pair, we are sure of alignment pair (seito , pupil).

Baseline

HM-BiTAM (Zhao and Xing, 08):

・HMM word alignment model + **topic model**

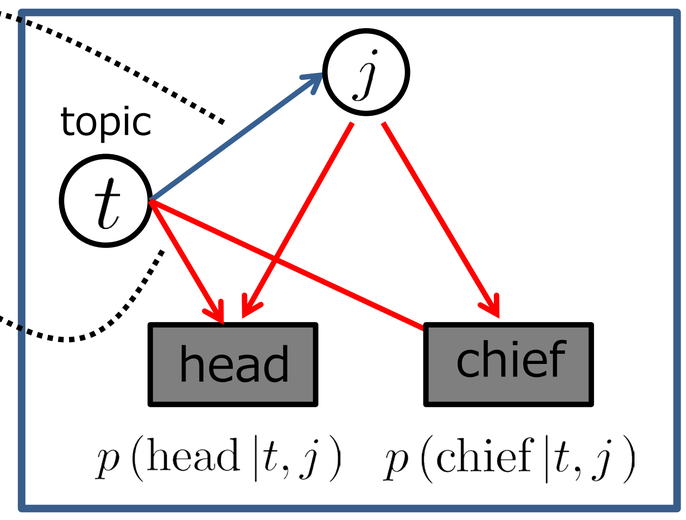
・Generative model of bilingual corpus $\{J, E\}$



$p(j|t)$

$p(e|t,j)$

common probabilistic model



Advantage of topic model:

・a word will have much less translation candidates due to constraints by the hidden topics.

➡ We obtain unambiguous word translation model.

Proposed

We propose a generative model of **synonym dictionary**

Note: synonym relations are context dependent

➡ We use a **topic model** to disambiguate the meaning of synonym pair

synonym dictionary

e	e′
head	chief
head	forefront
student	pupil
.....	.....

share **common** parameter sets Ψ and **jointly** train:

$$\arg\max_{\Psi} \{\log p(\mathbf{J}, \mathbf{E}; \Psi) + \zeta \log p(\{e, e'\}; \Psi)\}$$

HM-BiTAM | Synonym Pairs Model

Experiment

Hansards (English-French)

・347 test sentences

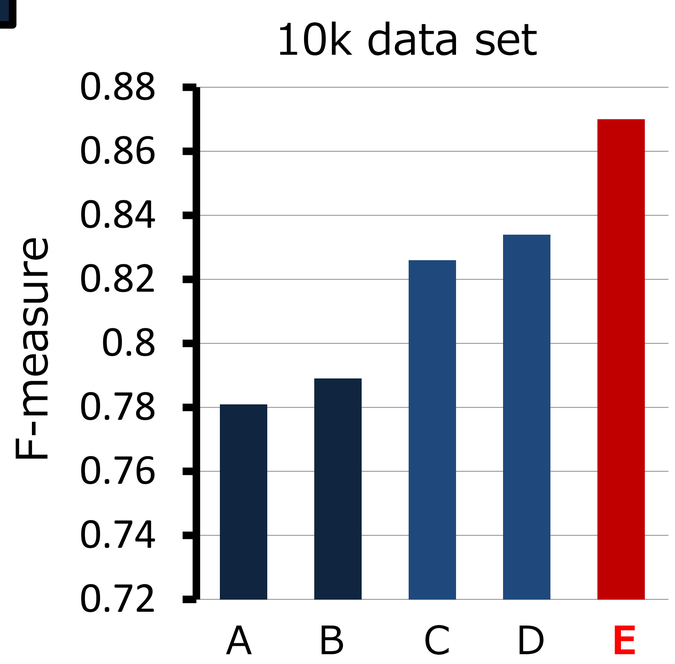
・100 development sentences

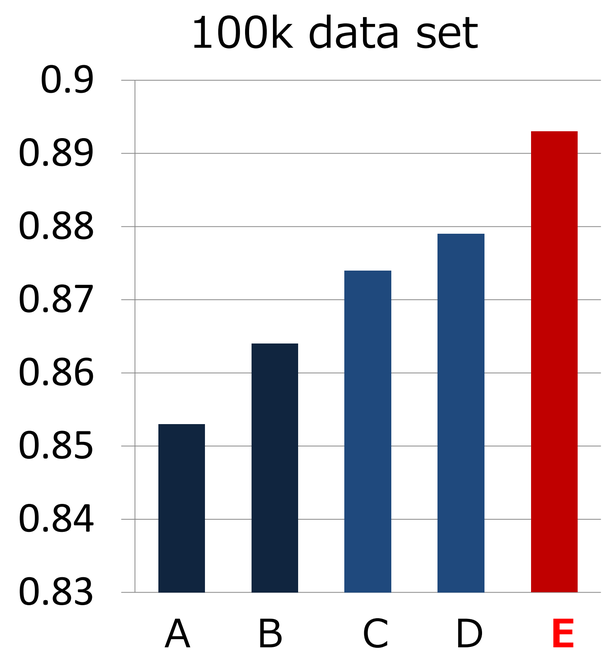
・randomly selected 10k and 100k training sentences

・obtained synonyms from WordNet

SRH: heuristics where all of the synonym pairs in the bilingual corpus are simply replaced with a representative word.

e.g. ~~head~~ ⟶ **chief**





A: GIZA++ (standard)
B: GIZA++ (SRH)
C: HM-BiTAM (standard)
D: HM-BiTAM (SRH)
E: Proposed

10k data set settings:
vocabularies (standard)
En 8578, Fr: 10791
vocabularies (SRH)
En: 5435, Fr: 9737
synonym pairs
En: 7756, Fr: 1677