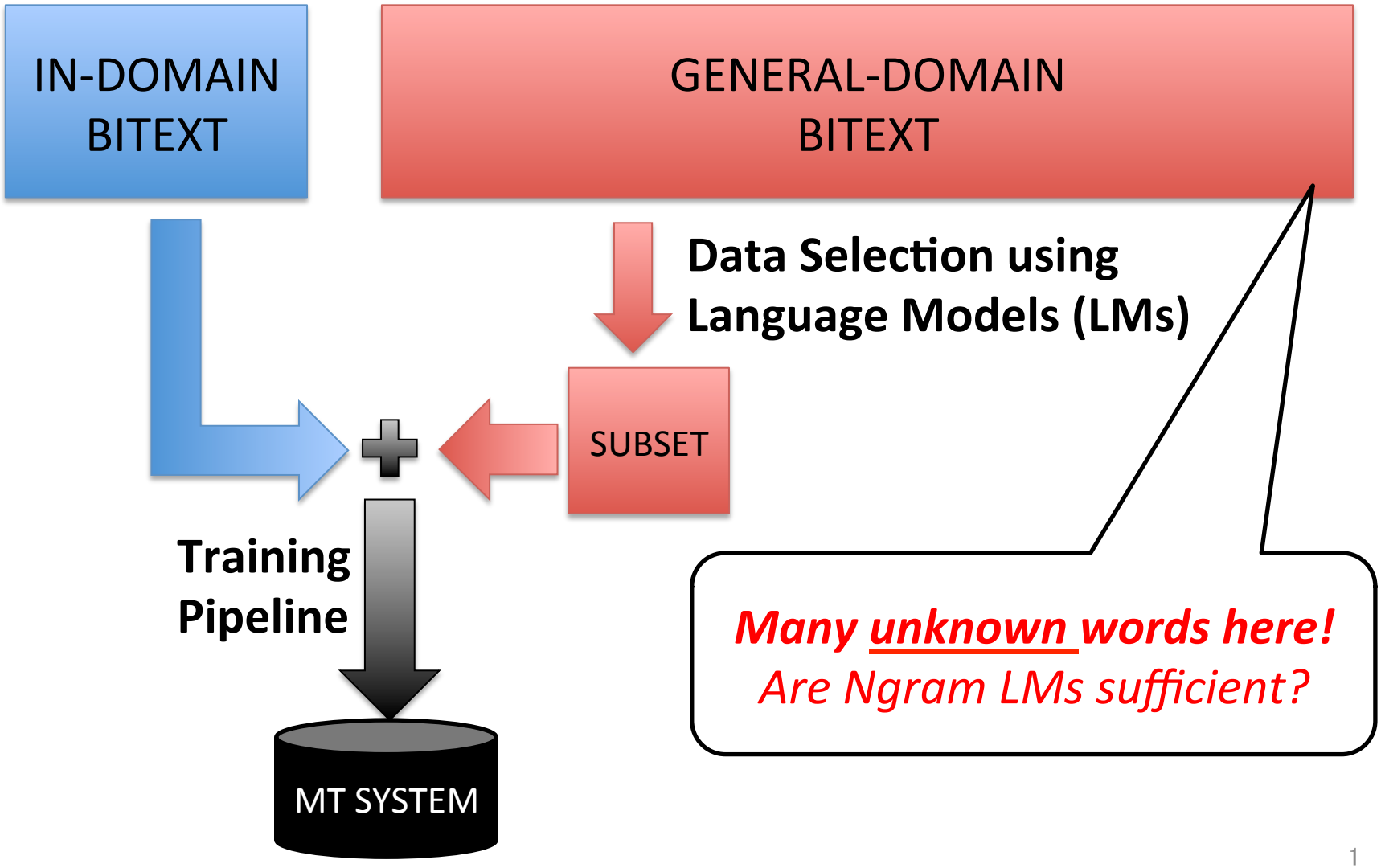
Adaptation Data Selection using Neural Language Models:
Experiments in Machine Translation

Kevin Duh
Graham Neubig

Katsuhito Sudoh
Hajime Tsukada

The Big Picture



Research Question:

- To investigate alternative LMs for data selection
- In particular: Neural LM
  - *Their continuous word representations have been shown to be robust to unknown words*

Result:

- Data selection by Neural LMs improve over Ngram LMs by 0.1-1.7 BLEU (TED Talks tasks).

Data Selection Criteria

[A. Axelrod, X. He, J. Gao. Domain Adaptation via Pseudo In-Domain Data Selection (EMNLP11)]

1. Score each General-Domain sentence-pair (e,f) by 4 LMs:

$$[\text{CrossEntropy}(LM^{e}_{IN}, e) - \text{CrossEntropy}(LM^{e}_{GENERAL}, e)]$$
$$+ [\text{CrossEntropy}(LM^{f}_{IN}, f) - \text{CrossEntropy}(LM^{f}_{GENERAL}, f)]$$

Prefer sentences similar to in-domain bitext

Prefer sentences dissimilar to average general domain bitext

2. Rank sentence pairs by score; threshold by validation set

Ngram vs. Recurrent Neural LM

$P(w_t|w_{t-1},w_{t-2})$

Backoff is needed for rare or unknown contexts (w_{t-1},w_{t-2})

e.g.

"recite Shakespeare's poem"

→P(poem|Shakespeare's recite)

"recite ScoobyDoo's poem"

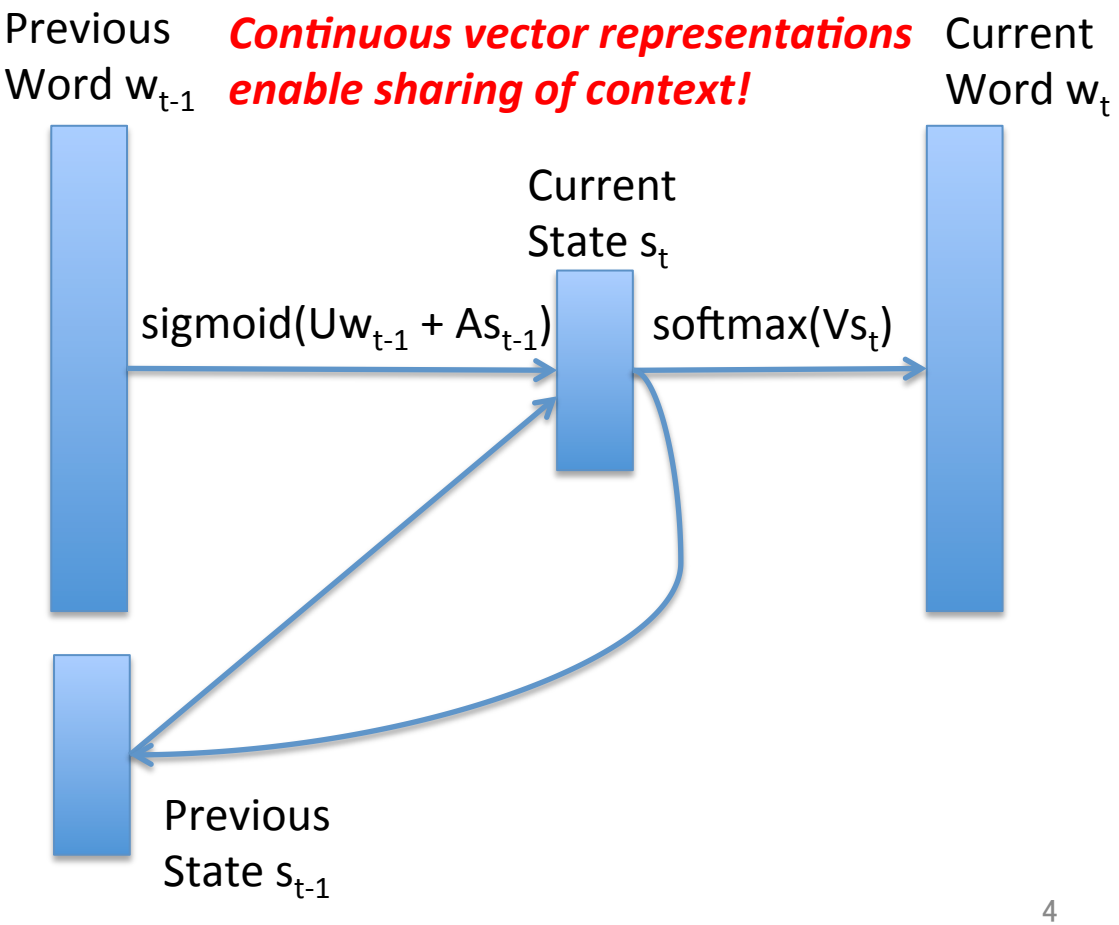
→P(poem|ScoobyDoo's recite)

→P(poem|ScoobyDoo's)

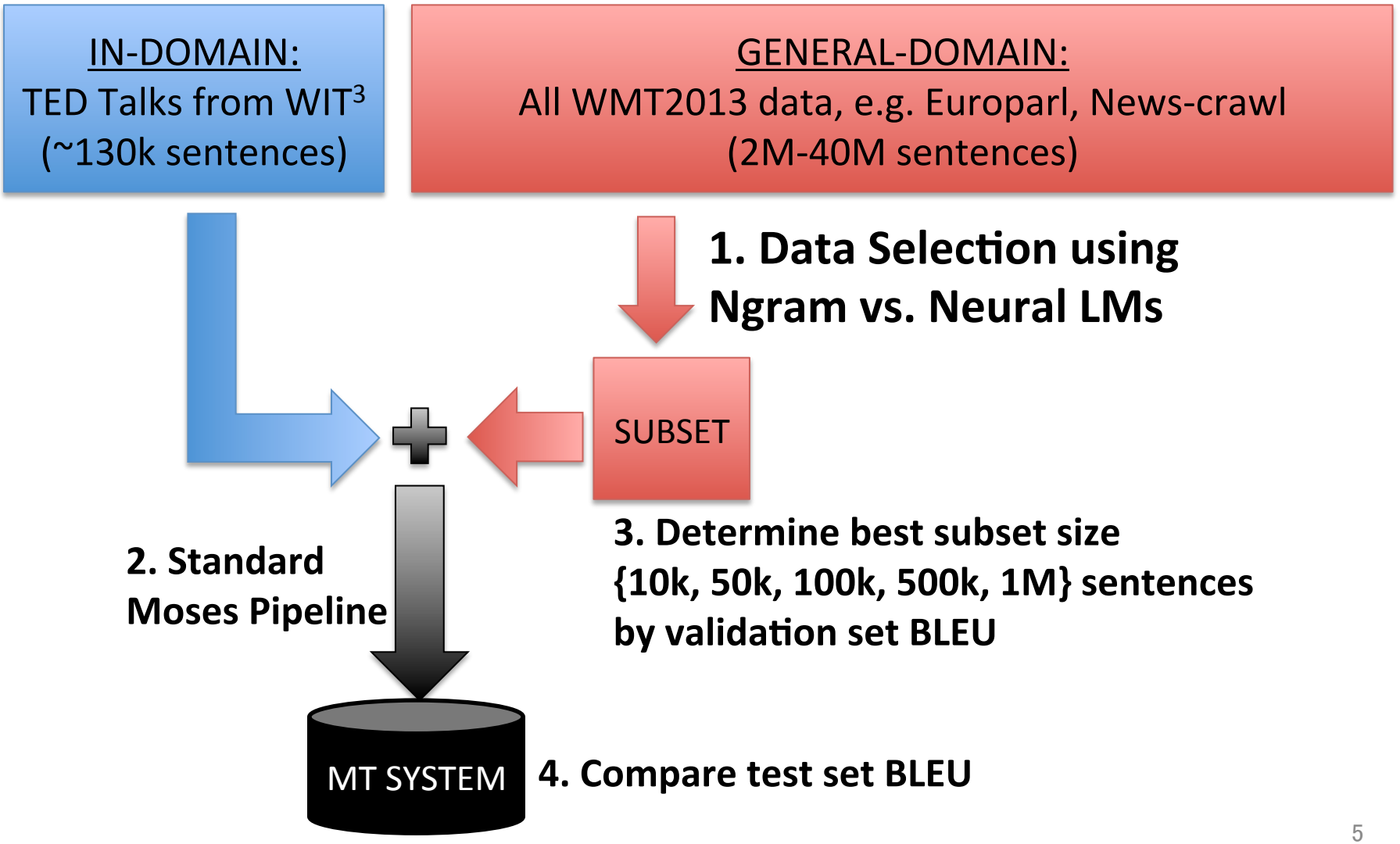
→P(poem)

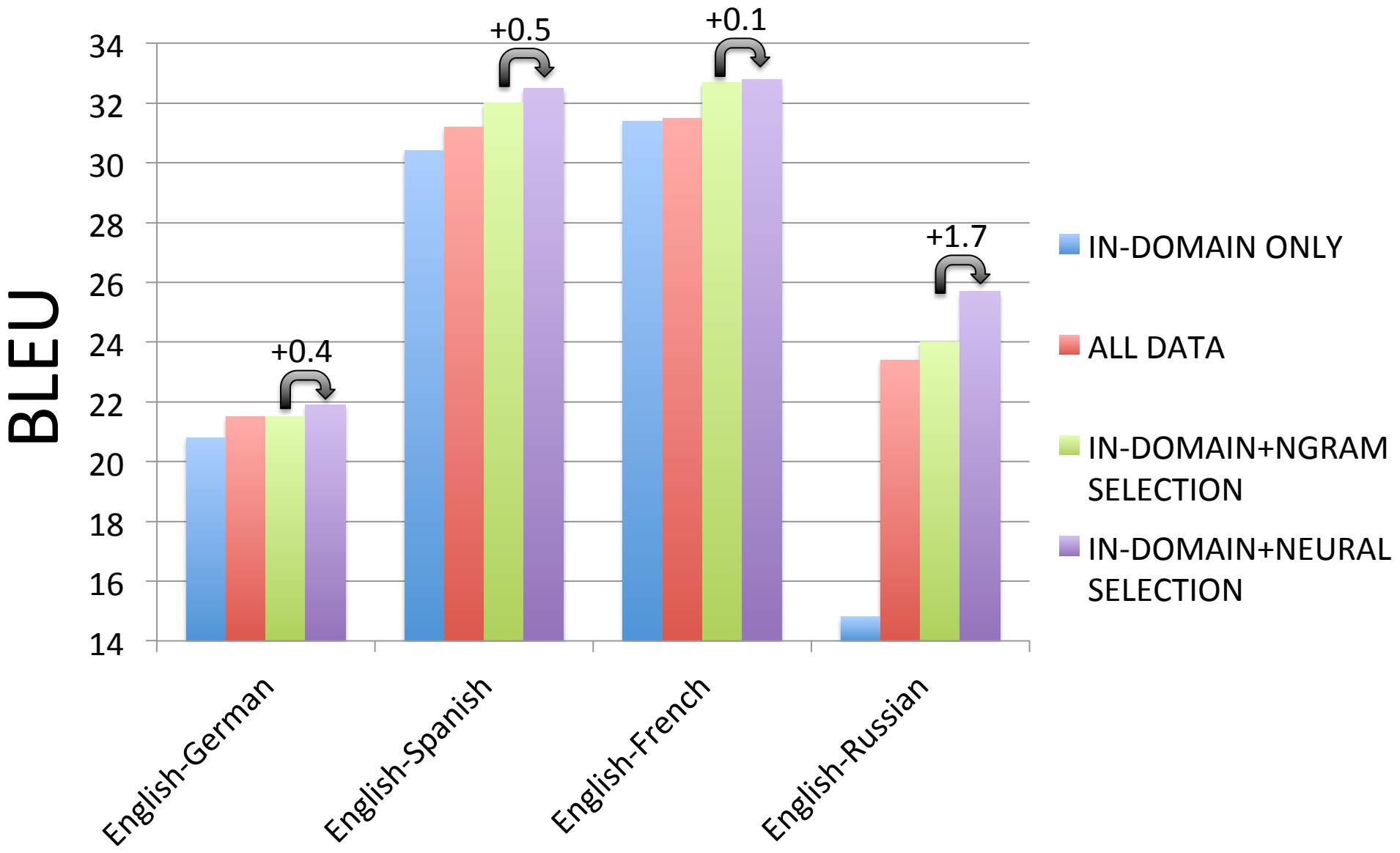
[T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, S. Khudanpur. Extensions of Recurrent Neural Net Language Model (ICASSP11)]

Continuous vector representations enable sharing of context!



Experimental Setup: 4 language-pairs




BLEU
34
32
30
28
26
24
22
20
18
16
14
+0.4
+0.5
+0.1
+1.7
English-German
English-Spanish
English-French
English-Russian
IN-DOMAIN ONLY
ALL DATA
IN-DOMAIN+NGRAM SELECTION
IN-DOMAIN+NEURAL SELECTION

Analysis

1. Are improvements due to lower OOV rate or better estimates of translation probability?
   - Force decoding gives similar BLEU → better estimates
2. How much overlap between sentences selected by Ngram vs. Neural LM?
   - 60-75% overlap, so this is incremental improvement
3. Computation time?

   - Fast to train Neural LMs for small in-domain set

Summary of paper in Haiku

(thanks to Chris Quirk for poetic inspiration)

These Neural LMs

Easy and good like N-grams

Why don't you try them?

Code/Scripts available: http://cl.naist.jp/~kevinduh/a/acl2013/