YJST at the NTCIR-12 MobileClick-2 Task

Yuya Ozawa

Taichi Yatsuka

Sumio Fujita

Agenda

- Introduction
- iUnit ranking subtask
  - Related works
  - Method
  - Experiments
  - Conclusion
  - Future work
- iUnit Summarization subtask
  - Related work
  - Method
  - Experiments
  - Conclusion
  - Future work

Introduction

- Mobile device
  - Limited display space of the devices
  - Web search services should be optimized to the mobile environments
- Motivations
  - Effectiveness of the language model based text matching
  - Effectiveness of dimension reduction approaches including distributional word representation
- Subtasks
  - Japanese iUnit Ranking subtask
  - Japanese iUnit Summarization subtask

iUnit Ranking Task

Related work

- Language Modeling in information retrieval
  - Language Modeling as a probabilistic distribution that captures statistical regularities of language generation.
  - In retrieval, document ranking according to the likelihood of generating the query based on each document model.
- Related approaches
  - Dirichlet prior smoothing: Zhai and Lafferty 2004.
  - Pitman-Yor smoothing: Momtazi and Klakow 2010.

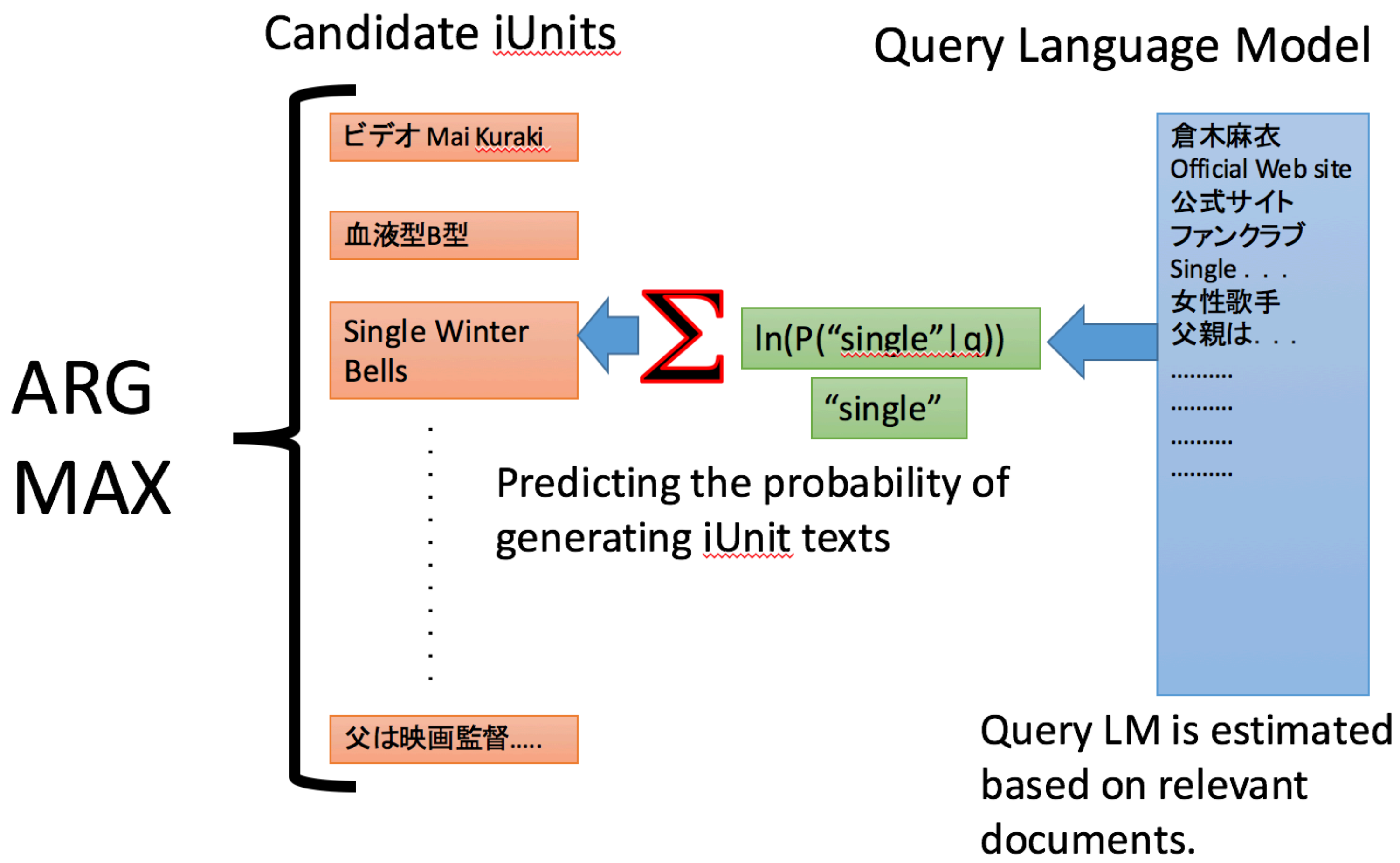
Japanese iUnit Ranking

METHOD

iUnit ranking subtask

- Language Modeling based approach
  - Score of each iUnit: probability of generating iUnit given a query language model
- Model
  - Query relevant Documents : $D = \{d_1, d_2, ..., d_n\}$
  - Document represented by word sets : $W = \{w_1, w_2, ..., w_n\}$
  - Query model : $P(w|q)$
  - Background (query non relevant) model : $P(w|o)$
- Data
  - Title and body in provided index data

Overview of Our Language Model



Dirichlet prior smoothing

- Uni-gram Dirichlet prior smoothing

$$P(w|q) = \frac{N_{D_q,w} + \mu P(w|o)}{N_{D_q} + \mu}$$

$$score(u,q) = \sum_{w \in W_u} \ln P(w|q)$$

- Bi-gram Dirichlet prior smoothing

$$P_{bi}(w_{i,i+1}|q) = \begin{cases} \frac{N_{D_q,w_{i,i+1}} + \mu P_{bi}(w_{i,i+1}|o)}{N_{D_q} + \mu} & (w_{i,i+1} \in D_q) \\ \lambda P(w_i|q) & (otherwise) \end{cases}$$

$$score(u,q) = \sum_{w_i, w_{i+1} \in W_u} \ln P_{bi}(w_{i,i+1}|q)$$

  - μ: hyper parameter
  - λ: down weighting factor

Other approaches

- KL Divergence

$$score(u,q) = D(P(w|q)||P(w|o)) = \sum_{w \in W_u} P(w|q) \ln \frac{P(w|q)}{P(w|o)}$$

- Pitman-Yor smoothing

$$P(w|q) = \frac{N_{D_q,w} - \delta + (\mu + \delta V_{D_q})P(w|o)}{N_{D_q} + \mu}$$

$$score(u,q) = \sum_{w \in W_u} \ln P(w|q)$$

  - V_{Dq}: vocabulary size in D_q
  - δ: hyper parameter

Japanese iUnit Ranking
EXPERIMENTS

Experiments

- Japanese iUnit Ranking task
- Training run results
  - Dirichlet prior smoothing and other approaches
- Test run results
  - Dirichlet prior smoothing

Training run results

Run description	Run detail	Q-Measure
Random ranking (ORG-R)	–	0.7201
Log Odds Ratio (ORG-L)	Laplace smth	0.7901
Vector Space Cosine	term freq	0.7715
Vector Space Cosine	Boolean	0.78
Vector Space+Background	Boolean	0.8003
Uni-gram Dirichlet priors	$\mu = 1, \alpha = 1$	0.8347
Uni-gram Dirichlet priors	$\mu = 0.5, \alpha = 1$	0.8352
Bi-gram Dirichlet priors	$\mu = 1, \alpha = 0$	**0.8399**
Mixture Dirichlet priors	$\mu = 1, \alpha = 0.5$	0.8375
KL-Divergence	Laplace smth	0.8108
Pitman-Yor	$\mu = 1, \delta = 0.1$	0.8321
iUnit LM	Dir prior $\mu = 1$	0.8258
iUnit LM+cotopic	Dir prior $\mu = 1$	0.8343
iUnit LM+coclick	Dir prior $\mu = 1$	0.8339
iUnit LM+cosession	Dir prior $\mu = 1$	0.8329
iUnit LM+chie	Dir prior $\mu = 1$	0.8345

Training run results

Run description	Run detail	Q-Measure
Random ranking (ORG-R)	–	0.7201
Log Odds Ratio (ORG-L)	Laplace smth	0.7901
Vector Space Cosine	term freq	0.7715
Vector Space Cosine	Boolean	0.78
Vector Space+Background	Boolean	0.8003
Uni-gram Dirichlet priors	$\mu = 1, \alpha = 1$	0.8347
Uni-gram Dirichlet priors	$\mu = 0.5, \alpha = 1$	0.8352
Bi-gram Dirichlet priors	$\mu = 1, \alpha = 0$	**0.8399**
Mixture Dirichlet priors	$\mu = 1, \alpha = 0.5$	0.8375
KL-Divergence	Laplace smth	0.8108
Pitman-Yor	$\mu = 1, \delta = 0.1$	0.8321
iUnit LM	Dir prior $\mu = 1$	0.8258
iUnit LM+cotopic	Dir prior $\mu = 1$	0.8343
iUnit LM+coclick	Dir prior $\mu = 1$	0.8339
iUnit LM+cosession	Dir prior $\mu = 1$	0.8329
iUnit LM+chie	Dir prior $\mu = 1$	0.8345

Test run results

Run description	Run detail	Q-Measure
Random Ranking (ORG-R)	–	0.7411
Log Odds Ratio (ORG-L)	Laplace smth	0.7269
Uni-gram Dirichlet priors	$\mu = 10, \alpha = 1$	**0.8072**
Bi-gram Dirichlet priors	$\mu = 1, \alpha = 0$	0.7965
Mixture Dirichlet priors	$\mu = 1, \alpha = 0.5$	0.8029
Uni-gram Dirichlet priors	$\mu = 0.5, \alpha = 1$	**0.8081**

Test run results

Run description	Run detail	Q-Measure
Random Ranking (ORG-R)	–	0.7411
Log Odds Ratio (ORG-L)	Laplace smth	0.7269
Uni-gram Dirichlet priors	$\mu = 10, \alpha = 1$	**0.8072**
Bi-gram Dirichlet priors	$\mu = 1, \alpha = 0$	0.7965
Mixture Dirichlet priors	$\mu = 1, \alpha = 0.5$	0.8029
Uni-gram Dirichlet priors	$\mu = 0.5, \alpha = 1$	**0.8081**

Conclusion

- We use Dirichlet prior smoothing in the LM-Based iUnit ranking approach
  - We carried out several experiments examining Uni-gram/Bi-gram iUnit/query language models
  - we achieved Q-score of 0.807 in a test run using a Uni-gram model

Future work

- Our approach only uses the divergence between query and background language models
- Adopt supervised learning to rank iUnits using several features including:
  - textual
  - nontextual

iUnit Summarization Task

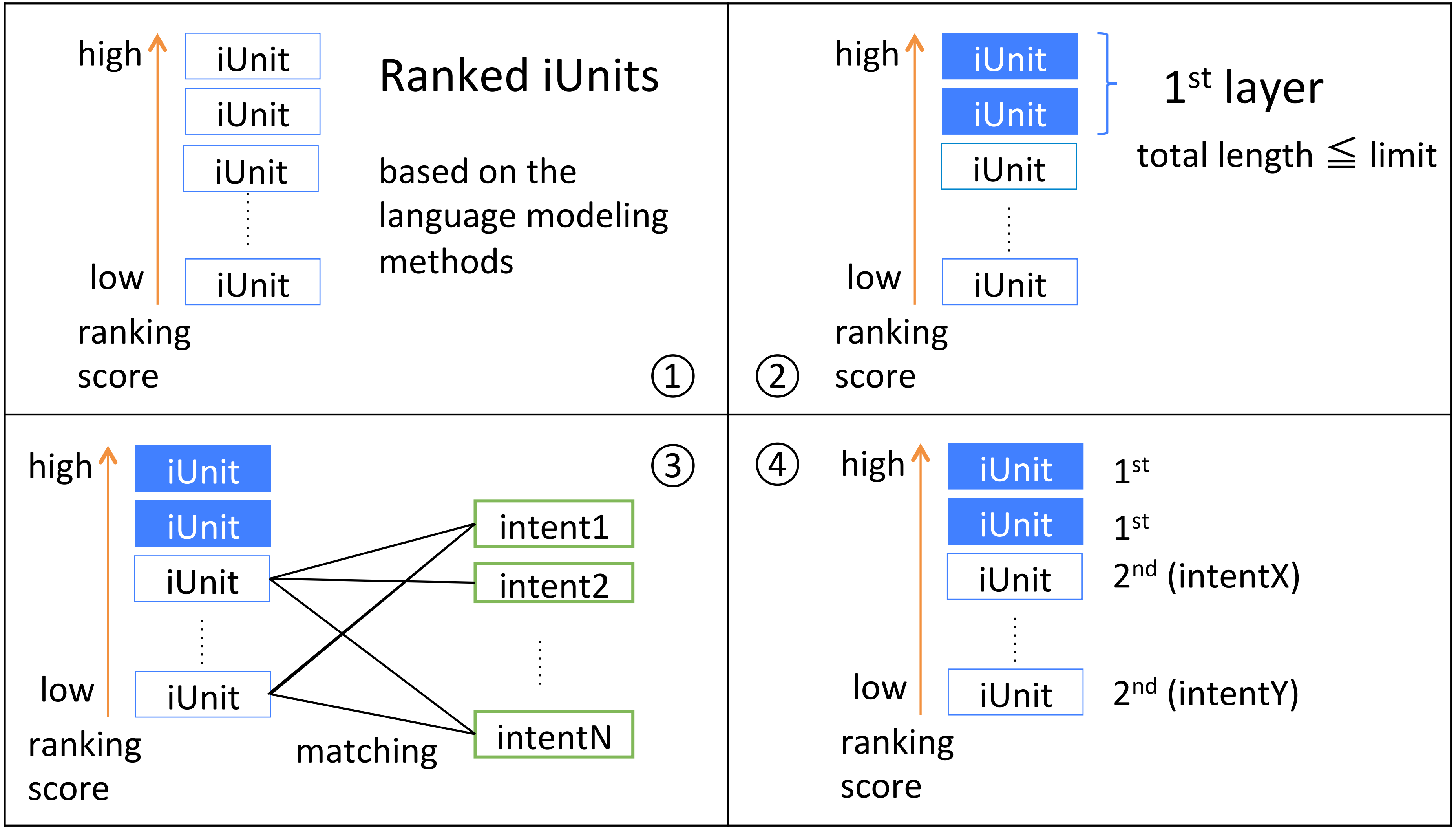
Related Work

- Efficient estimation of word representations in vector space. (T. Mikolov et al., ICLR 2013)

Japanese iUnit Summarization

METHOD

LM-based Two-layer iUnit Summarization Baseline



LM-based Two-layer iUnit Summarization Baseline

- The computation of the iUnit score against each second layer intent as follows:

$$Score(u, i) \quad = \quad R(u) \cdot Sim(u, i)$$

  - *u , i* : iUnit and intent
  - *R(u)* : the iUnit ranking score from the ranking method
  - *Sim(u, i)* : the score of intent matching

Set based Intent Matching

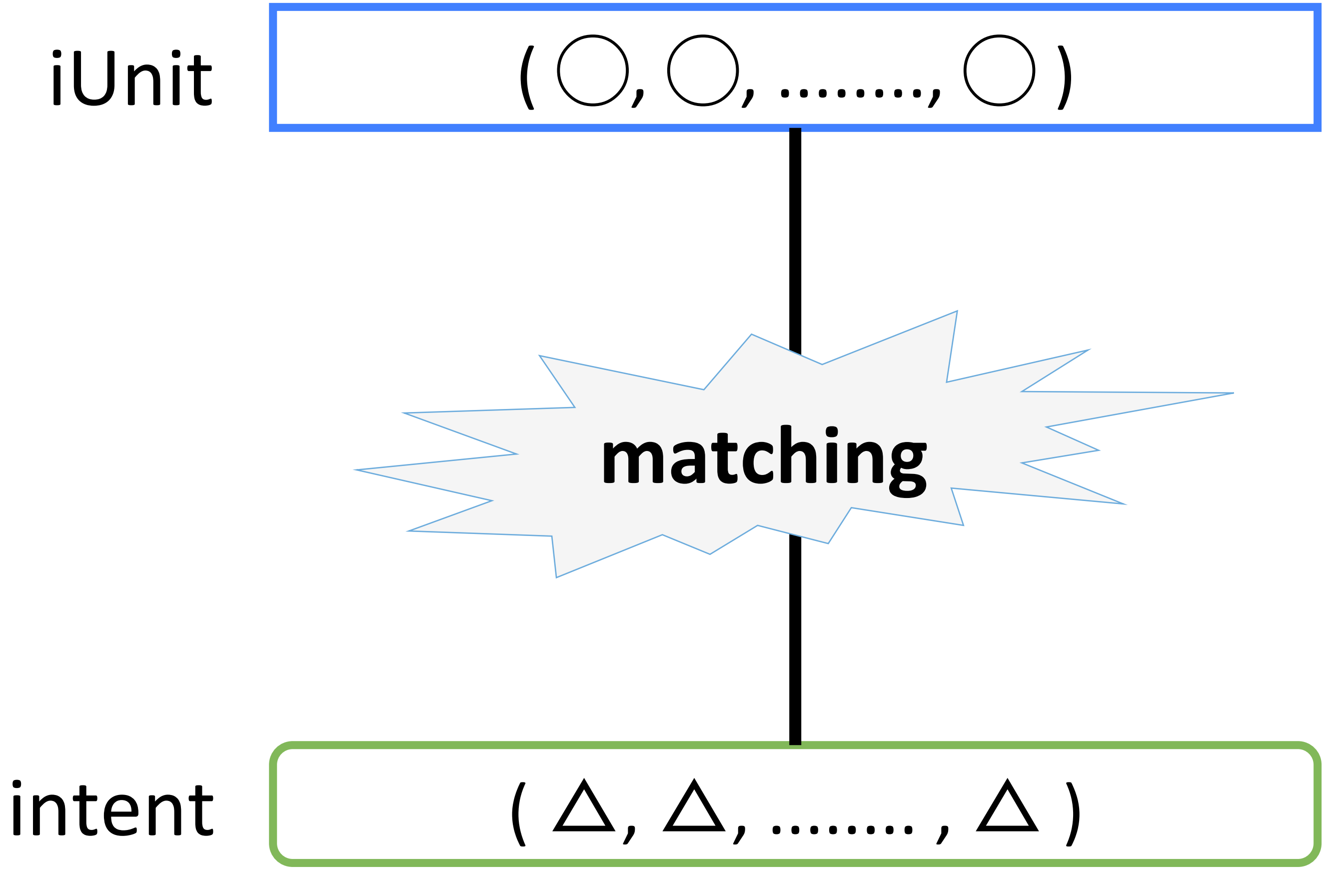
- asymmetric similarity function (organizer's baseline)

$$Sim_{set}(u, i) \quad = \quad \frac{|W_u \cap W_i|}{|W_i|}$$

  - W_x : the set of words contained in x

※ $Sim_{set}(u, i)$ becomes *0* when there is no common word between u and i.

Word Embedding based Intent Matching



Word Embedding based Intent Matching

- iUnit embedding : *Embu*

$$Emb_u = \sum_{w_u \in W_u} Emb_{w_u}$$

  - Emb_{w_u} : embedding of the word w_u.

- intent embedding : *Embi*

$$Emb_i = \sum_{w_i \in W_i} Emb_{w_i}$$

  - Emb_{w_i} : embedding of the word w_i

Word Embedding based Intent Matching

- Simlirality calculation :

$$Sim_{emb}(u,i) \quad = \quad \cos(Emb_u, Emb_i)$$

  - *cos(X,Y)* is cosine similarity.

  - In Additional experiments
    - We also tried another similarity measure based on the Euclidean distance between vectors.

Japanese iUnit Summarization

EXPERIMENTS

Parameters

- embedding training parameter
  - data : given HTML's <body> without tag
  - vector size : 200
  - model : CBoW
  - window size : 5
  - implimentation : Google Code Archive word2vec
    - https://code.google.com/archive/p/word2vec/

Results

Submit #	Run type	Ranking	Intent Matching	Limit	M-measure
123	ORG-T	Log Odds Ratio LM	Set based	280	17.4376
437	Addition	Log Odds Ratio LM	Emb+Cos	280	19.094
131	Official	KL-Div LM	Set based	280	21.0259
173	Official	Dir priors LM	Emb+Cos	280	25.8498
231	Official	Dir priors LM	Emb+Cos	0	13.9927
324	Official	Dir priors LM	Emb+Cos	252	25.6084
419	Addition	Dir priors LM	Set based	280	26.7036
442	Addition	Dir priors LM	Emb+Euclidean	280	26.6096

- **Limit** indicates the first layer length limit.

Results

Submit #	Run type	Ranking	Intent Matching	Limit	M-measure
123	ORG-T	Log Odds Ratio LM	Set based	280	17.4376
437	Addition	Log Odds Ratio LM	Emb+Cos	280	19.094
131	Official	KL-Div LM	Set based	280	21.0259
173	Official	Dir priors LM	Emb+Cos	280	25.8498
231	Official	Dir priors LM	Emb+Cos	0	13.9927
324	Official	Dir priors LM	Emb+Cos	252	25.6084
419	Addition	Dir priors LM	Set based	280	26.7036
442	Addition	Dir priors LM	Emb+Euclidean	280	26.6096

- **Limit** indicates the first layer length limit.

Results

Submit #	Run type	Ranking	Intent Matching	Limit	M-measure
123	ORG-T	Log Odds Ratio LM	Set based	280	17.4376
437	Addition	Log Odds Ratio LM	Emb+Cos	280	19.094
131	Official	KL-Div LM	Set based	280	21.0259
173	Official	Dir priors LM	Emb+Cos	280	25.8498
231	Official	Dir priors LM	Emb+Cos	0	13.9927
324	Official	Dir priors LM	Emb+Cos	252	25.6084
419	Addition	Dir priors LM	Set based	280	26.7036
442	Addition	Dir priors LM	Emb+Euclidean	280	26.6096

- **Limit** indicates the first layer length limit.

Results

- Compare Embedding+Cosine method(#173) and set based method(#419) by query basis
  - 23 queries : #173 performed better
  - 4 queries : performed equally
  - 73 queries : #419 performed better

Results

Submit #	Run type	Ranking	Intent Matching	Limit	M-measure
123	ORG-T	Log Odds Ratio LM	Set based	280	17.4376
437	Addition	Log Odds Ratio LM	Emb+Cos	280	19.094
131	Official	KL-Div LM	Set based	280	21.0259
173	Official	Dir priors LM	Emb+Cos	280	25.8498
231	Official	Dir priors LM	Emb+Cos	0	13.9927
324	Official	Dir priors LM	Emb+Cos	252	25.6084
419	Addition	Dir priors LM	Set based	280	26.7036
442	Addition	Dir priors LM	Emb+Euclidean	280	26.6096

- **Limit** indicates the rst layer length limit.

Results

Submit #	Run type	Ranking	Intent Matching	Limit	M-measure
123	ORG-T	Log Odds Ratio LM	Set based	280	17.4376
437	Addition	Log Odds Ratio LM	Emb+Cos	280	19.094
131	Official	KL-Div LM	Set based	280	21.0259
173	Official	Dir priors LM	Emb+Cos	280	25.8498
231	Official	Dir priors LM	Emb+Cos	0	13.9927
324	Official	Dir priors LM	Emb+Cos	252	25.6084
419	Addition	Dir priors LM	Set based	280	26.7036
442	Addition	Dir priors LM	Emb+Euclidean	280	26.6096

- **Limit** indicates the rst layer length limit.

Results

- Compare Euclidean distance method(#442) and set based method(#419) by query basis
  - 44 queries : #442 performed better
  - 13 queries : performed equally
  - 43 queries : #419 performed better

- The vector similarity measure greatly affects the effectiveness of intent matching of word embedding based.
- This suggests that the better usage of word embedding representation leads to more effective intent matching solutions.

Results

Submit #	Run type	Ranking	Intent Matching	Limit	M-measure
123	ORG-T	Log Odds Ratio LM	Set based	280	17.4376
437	Addition	Log Odds Ratio LM	Emb+Cos	280	19.094
131	Official	KL-Div LM	Set based	280	21.0259
173	Official	Dir priors LM	Emb+Cos	280	25.8498
231	Official	Dir priors LM	Emb+Cos	0	13.9927
324	Official	Dir priors LM	Emb+Cos	252	25.6084
419	Addition	Dir priors LM	Set based	280	26.7036
442	Addition	Dir priors LM	Emb+Euclidean	280	26.6096

- **Limit** indicates the rst layer length limit.

Conclusion

- We adopted a new intent matching method using word embedding representations.
  - This leads to a finer allocation of relevant iUnits to subtly related intents in the second layer.
- We achieved M-measure of **25.8498**.
  - the **best** of official runs of the Japanese iUnit Summarization Subtask
- Additional experiments suggest the possibility of further improvements.
  - with more effective similarity matching

Future Work

- Examining better word embedding representations
- Examining other similarity measures to vectorial matching
    - KL-divergence, Jaccard coefficient and so on
- Optimizing the strategy in view of M-measure

EOP