Overview of the NTCIR-10 Cross-lingual Link Discovery Task

Introduction

The NTCIR-10 Cross-lingual Link Discovery Task

- Shortly, the CrossLink-2 task
- The CrossLink-2 task @NTCIR-10 is the second round of Cross-Lingual Link Discovery evaluation.
- New tasks, new settings, new challenges, new participants.

Why CLLD is important?

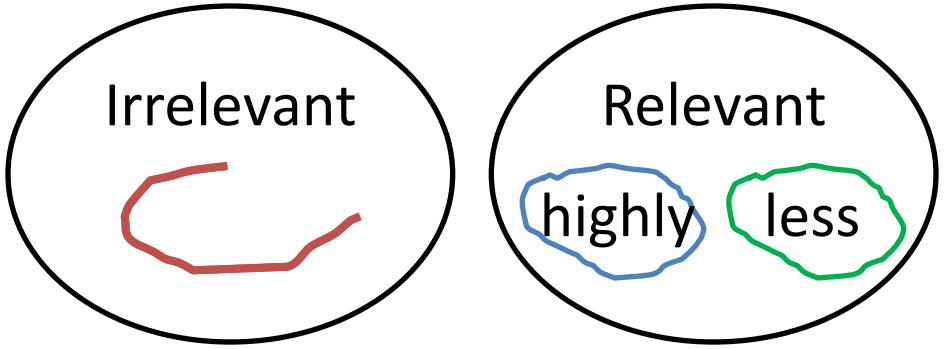
布丁 [编辑]

维基百科，自由的百科全书

布丁一詞源自音譯英語的Pudding，意譯則為「奶凍」。廣義來說泛指由漿狀的材料凝固成固體果凍狀的食品，如聖誕布丁、麵包布丁等，常見製法包括焗及蒸等。狹義來說，布丁是一種冷凍的甜品，主要材料為雞蛋，類似果凍。

聖誕布丁

市售大量生產的布丁，為了節省成本，故極少使用雞蛋與牛奶當原料，而是使用「布丁粉」沖泡後冷藏。其主要原料為海藻抽出物、食用色素等。



It is all about efficient and effective information and knowledge discovery in a multi-lingual environment (e.g. Wikipedia with over 10,000,000 articles but rarely cross-lingual linked).

CrossLink-1 Review (1)

- CLLD is a way of providing easy access to the cross-lingual information and break the language barrier, and it is concerned with automatically finding potential links between documents in different languages.
- English to CJK language tasks:
  - English to Chinese CLLD (E2C)
  - English to Japanese CLLD (E2J)
  - English to Korean CLLD (E2K)

CrossLink-1 Review (2)

- The goal of the CrossLink task to create a reusable resource for evaluating automated cross language link discovery approaches. The results of this research will be used in building and refining systems for automated link discovery.
- With the developed evaluation framework @NTCIR-9, we identified many good CLLD approaches.

CrossLink-2 Tasks

- New tasks:
  - Chinese to English CLLD (C2E)
  - Japanese to English CLLD (J2E)
  - Korean to English CLLD (K2E)

- Participants will have to deal with extra NLP problems such as text segmentation when trying to cross link documents as there are no word boundaries in Chinese / Japanese text, and in Korean *eojeol*.

Possible problems

- Will natural language processing be a problem?
- Will a same method used previous in the E2CJK task still work on the CJK2E tasks?
- …

Example: Chinese Segmentation

Such as Chinese segmentation:

- 胸甲骑兵在腓特烈大帝和拿破仑的军队**中都**扮演过非常重要的角色。(Cuirassier plays a very important role in the armies of Friedrich II von Preußen and Napoléon Bonaparte)

中 and 都 mean “in, middle, ...” and “both, city, ..” separately, but together they (中都) are often used as place names (e.g. an old name for *Beijing* city).

- 胸甲骑兵放弃了对躯干**部分和**腿部的严密防护 (Cuirassier gives up the protection for part of body and legs)

Without proper segmentation the two words 部分 (means “*part*”) 和 (means “*and*”) in the second sentence could be easily processed as one and linked to the less relevant mathematical article -部分和 (Series (mathematics)), .

Experiment

New Collections

- A complete new set of CEJK Wikipedia document collections which were built from recent Wikipedia database dumps (2012).

LANG	#DOC	SIZE	DATE OF DUMP
Chinese	404,620	3.6GB	11/01/2012
English	3,581,772	33.0GB	04/01/2012
Japanese	858,610	9.8GB	04/01/2012
Korean	297,913	2.2GB	22/01/2012

New Participants

GROUP	ORGANISATION
DCU	Dublin City University
III	Institute for Information Industry
KECIR	Shenyang Aerospace University
KMI	The Open University
KSLP	Kyungsung University
NTHU	National Tsing Hua University
OKSAT	Osaka Kyoiku University
QUT	Queensland University of Technology
RDLL	Ritsumeikan University
UKP	TU Darmstadt

The list of participant teams @NCIR-10 CrossLink-2

Submissions

GROUP	CJK2E			E2CJK		
	C2E	J2E	K2E	E2C	E2J	E2K
DCU	2	0	0	2	0	0
III	3	0	0	1	0	0
KECIR	4	0	0	0	0	0
KMI	3	3	3	2	2	2
KSLP	0	0	1	0	0	0
NTHU	3	1	0	0	0	0
OKSAT	2	2	2	2	2	2
QUT	2	2	2	1	1	1
RDLL	0	5	0	0	0	0
UKP	3	3	3	0	0	0
Sub-total	22	16	11	8	5	5
Total	49			18		

In total, there were 10 teams who submitted 49 (CJK2E) + 18 (E2CJK) runs

Anchor identification is hard

Basically, it means:

- Technically, it looks simple but specifying anchor offset correctly is not easy.
  - lots of submitted runs contain incorrect anchors.
  - Incorrect anchors mean they will not be pooled
- What should a meaningful and relevant anchor?
  - To many to choose, e.g. n-grams, phrases
  - How to rank the anchors?

Assessment and Evaluation

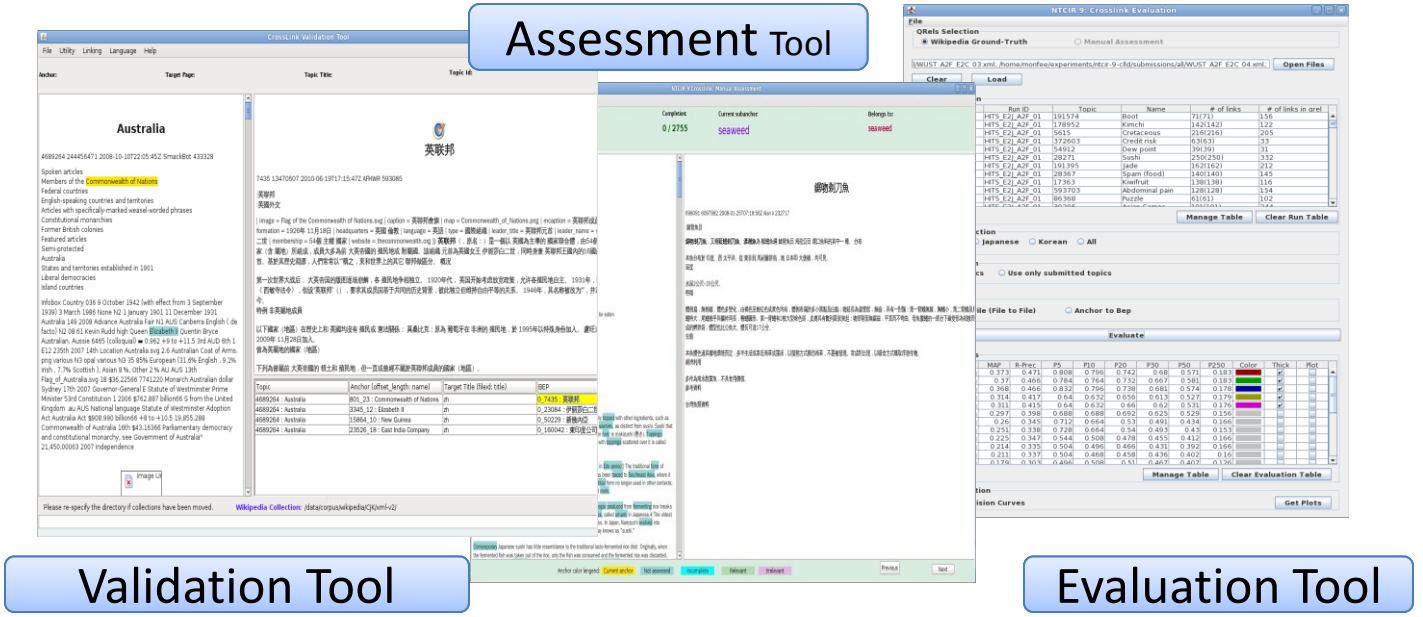
Assessment is hard too

- Three different language subtasks (E2CJK and CJK2E)
- Two set of links (submission and Wikipedia ground-truth)
- Hard to find appropriate assessors
- Many link to be assessed

Submitted Links			**Wikipedia Ground Truth Links**		
Task	#Total	#Average	Task	#Total	#Average
en-ja	24779	991	en-ja	1913	77
en-ko	22143	886	en-ko	1033	41
en-zh	23142	926	en-zh	1343	54
ja-en	34392	1376	ja-en	1890	76
ko-en	33179	1327	ko-en	1200	48
zh-en	84627	3385	zh-en	1478	59

Same Evaluation Framework

- Same evaluation framework but with new settings
- Same evaluation metrics: LMAP, R-Prec, P@N
- Assessment Types: Automatic (Wikipedia Ground Truth), Manual (human in the loop)



Three evaluation scenarios

- Overall, we have two types of tasks (E2CJK, CJK2E) with three evaluation scenarios:
  - F2F evaluation with Wikipedia ground-truth (GT F2F)
  - F2F evaluation with manual assessment results (MA F2F)
  - A2F evaluation with manual assessment results (MA A2F)

Precision and Recall – *F2F*

$$\text{Precision}_{f2f} = \frac{\text{number of correct links}}{\text{number of identified links}}$$

$$\text{Recall}_{f2f} = \frac{\text{number of correct links}}{\text{number of links in qrel}}$$

Precision and Recall – *A2F*

$$f_{anchor}(i) = \begin{cases} 1, & \text{if relevant with} \geq 1 \text{ relevant targe} \\ 0, & \text{otherwise} \end{cases}$$

$$f_{link}(j) = \begin{cases} 1, & \text{if relevant} \\ 0, & \text{otherwise} \end{cases}$$

$$Precision_{a2f} = \left(\sum_{i=1}^{n} \left(f_{anchor}(i) \right) \times \frac{\sum_{j=1}^{k_i} f_{link}(j)}{k_i} \right) / n$$

$$Recall_{a2f} = \left(\sum^{n} \left(f_{anchor}(i) \right) \times \frac{\sum_{j=1}^{k_i} f_{link}(j)}{k} \right) / N$$

System Evaluation Metrics

- $LMAP = (\sum_{t=1}^{n} \frac{\sum_{k=1}^{m} p_{kt}}{m})/n$
- $R - Prec = \sum_{t=1}^{n} P_t \text{ @ } R \,/n$
- *Precision-at-N* is computed using the average precision for all topics (source articles) at a pre-defined position *N* in the results list. Values of *N* were chosen as: 5, 10, 20, 30, 50, and 250.

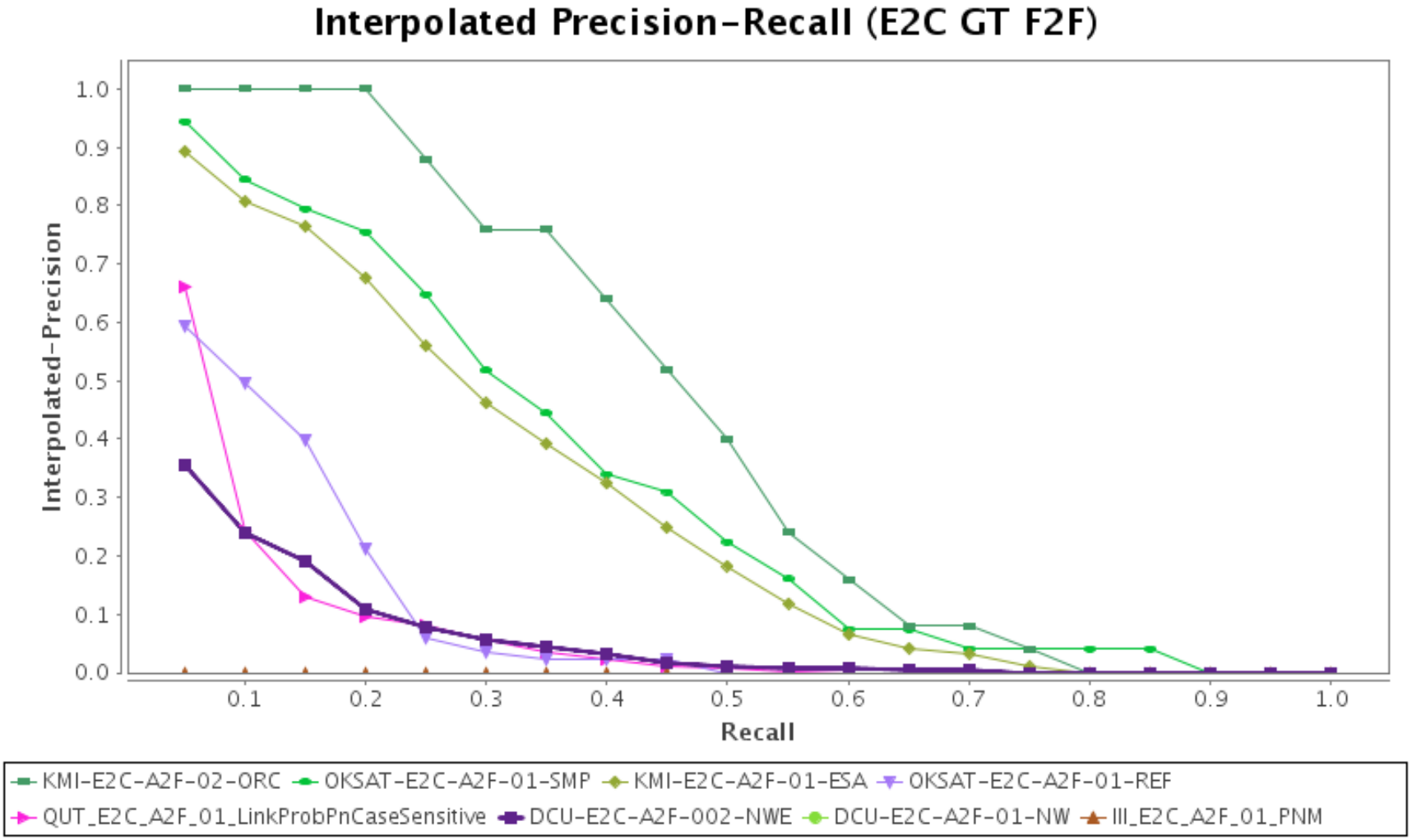
Evaluation Results

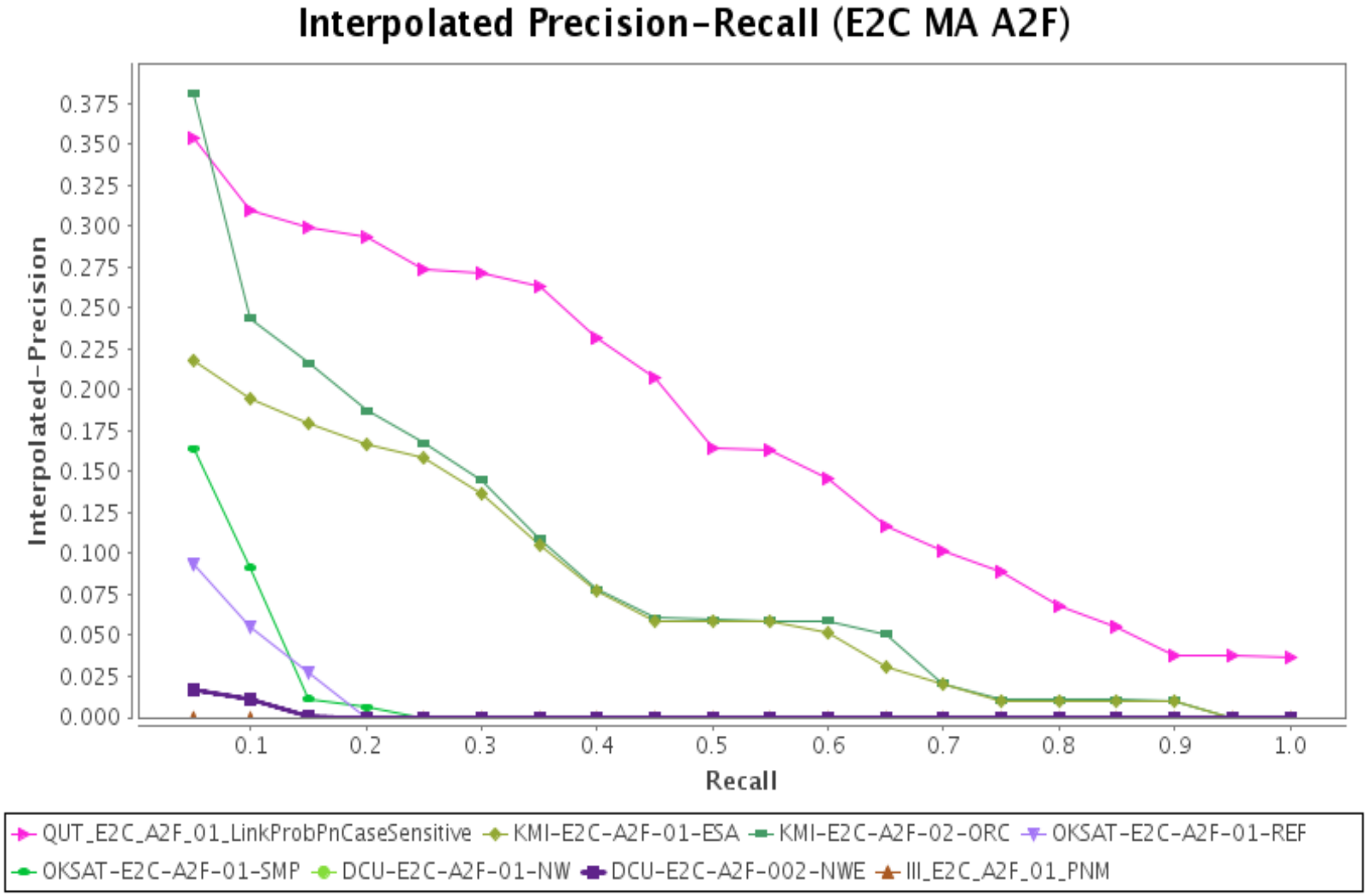
Evaluation Results – E2CJK

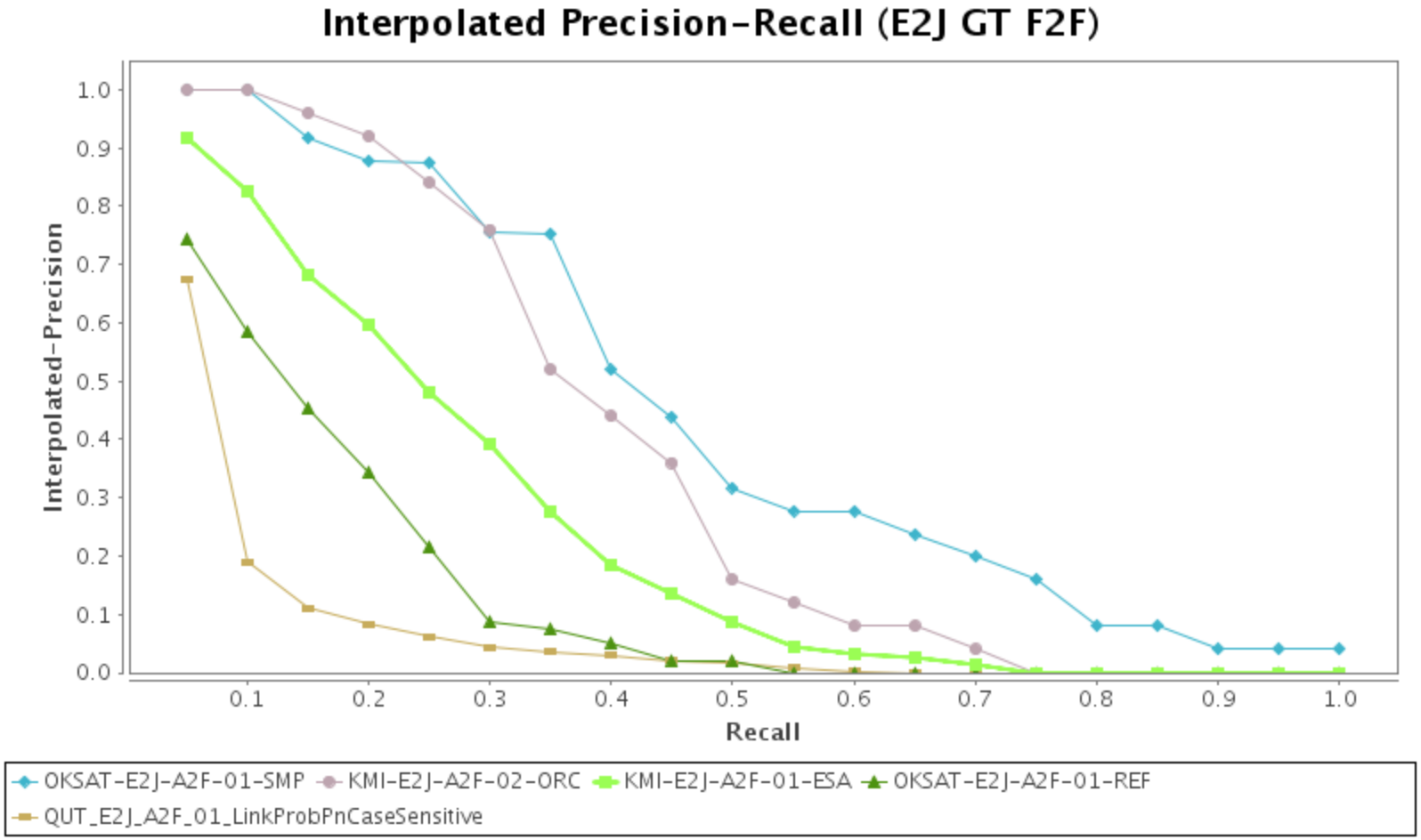
F2F GT	F2F MA	A2F MA
English-to-Chinese ***LMAP:*** **KMI**, OKSAT, QUT ***Precision-at-5:*** KMI, OKSAT, QUT English-to-Japanese ***LMAP:*** **OKSAT,** KMI, QUT ***Precision-at-5:*** **KMI**, OKSAT, QUT *English-to-Korean* ***LMAP:*** **OKSAT**, KMI, QUT ***Precision-at-5:*** **OKSAT**, KMI, QUT	English-to-Chinese ***LMAP:*** **KMI**, QUT, OKSAT ***Precision-at-5:*** **KMI**, QUT, OKSAT English-to-Japanese ***LMAP:*** **KMI**, OKSAT, QUT ***Precision-at-5:*** KMI, OKSAT, QUT English-to-Korean ***LMAP:*** **KMI**, OKSAT, QUT ***Precision-at-5:*** KMI, OKSAT, QUT	English-to-Chinese ***LMAP:*** **QUT**, KMI, OKSAT ***Precision-at-5:*** **KMI**, QUT, OKSAT English-to-Japanese ***LMAP:*** **KMI**, OKSAT, QUT ***Precision-at-5:*** **KMI**, OKSAT, QUT English-to-Korean ***LMAP:*** **KMI**, OKSAT, QUT ***Precision-at-5:*** **KMI**, OKSAT, QUT

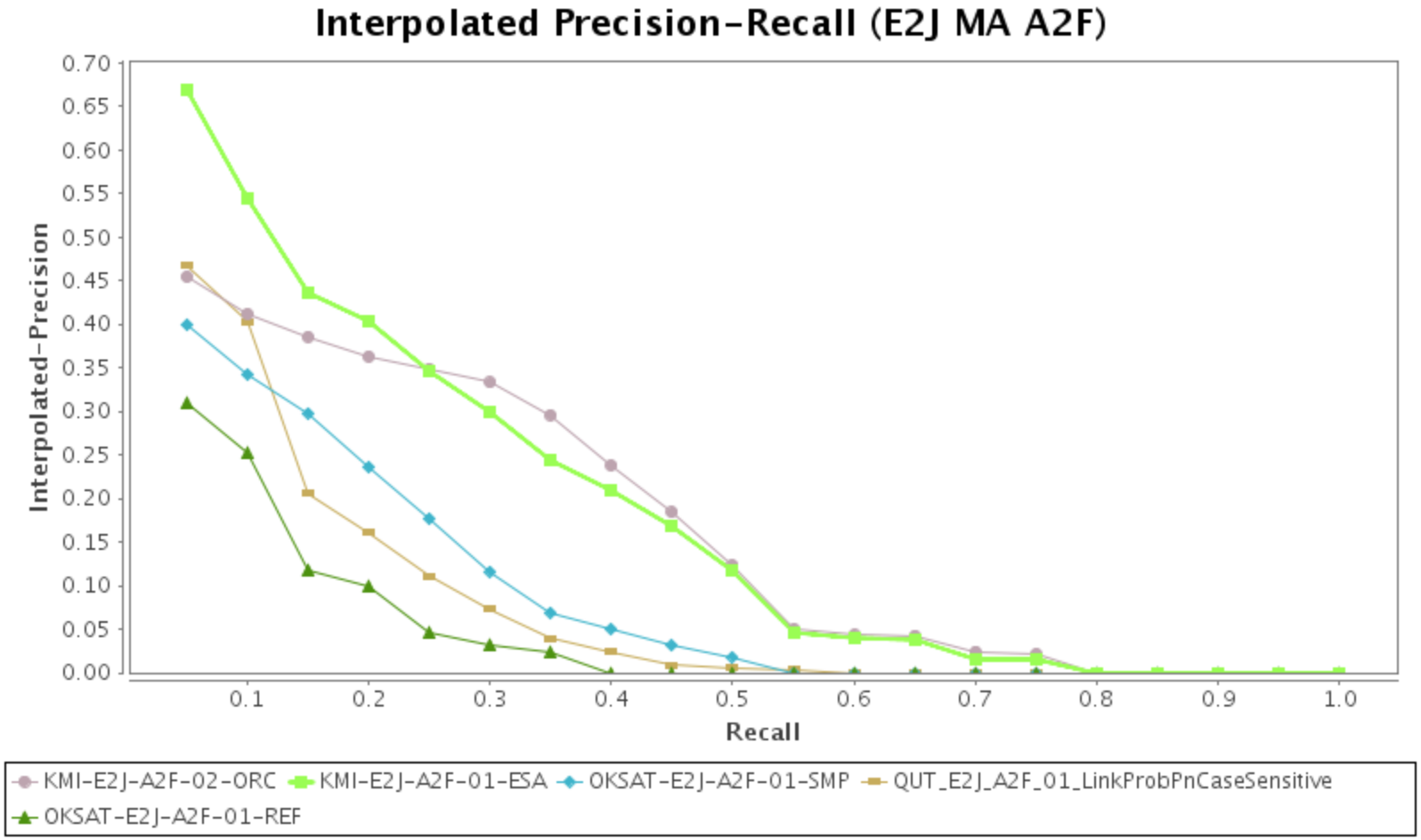
Evaluation Results – CJK2E

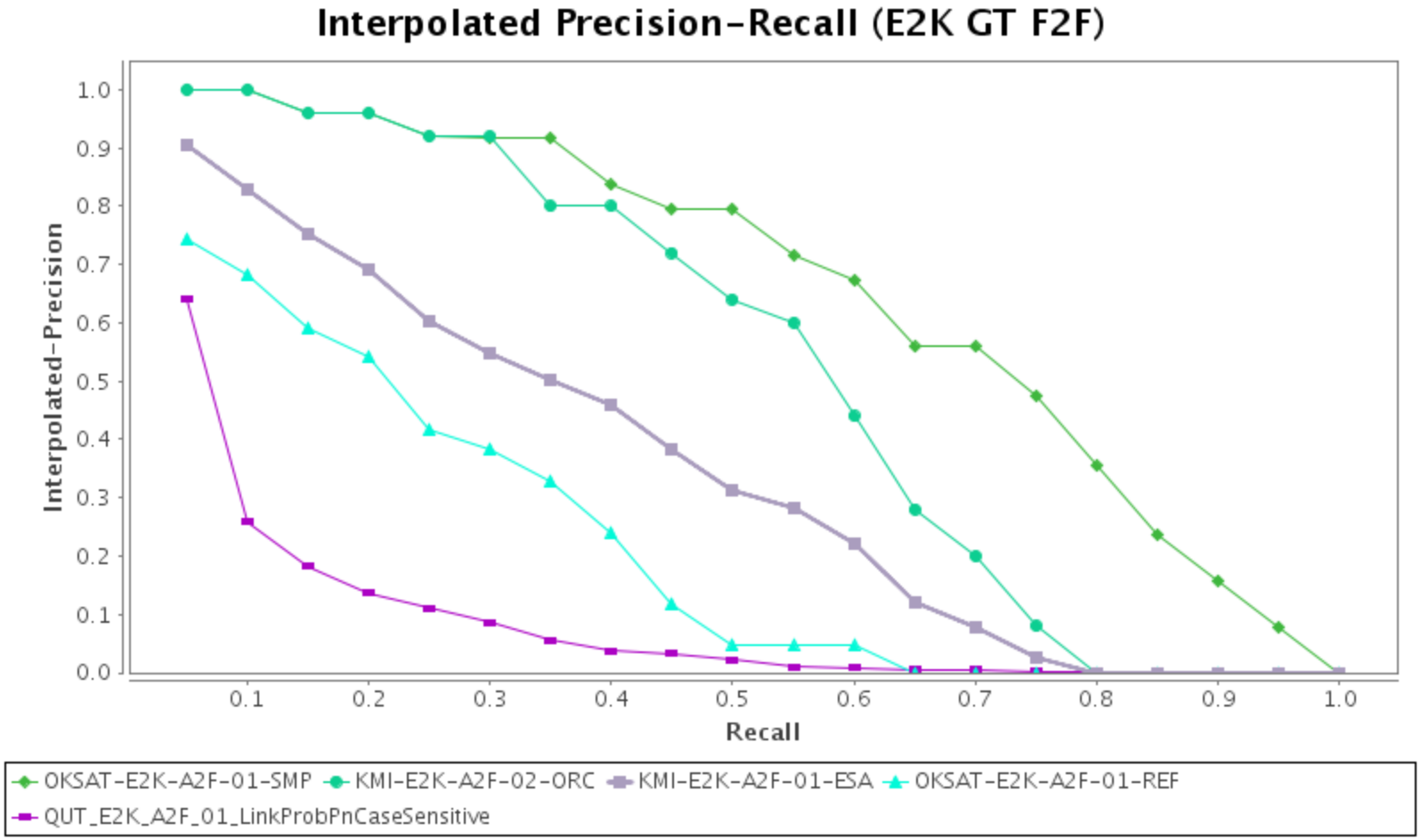
F2F GT	F2F MA	A2F MA
English-to-Chinese ***LMAP:*** **OKSAT**, KMI, UKP ***Precision-at-5:*** **OKSAT**, UKP, KMI English-to-Japanese ***LMAP:*** **OKSAT**, KMI, UKP ***Precision-at-5:*** **KMI**, OKSAT, UKP *English-to-Korean* ***LMAP:*** **OKSAT**, KSLP, KMI ***Precision-at-5:*** **OKSAT**, KSLP, KMI	English-to-Chinese ***LMAP:*** **QUT**, KMI, OKSAT ***Precision-at-5:*** **OKSAT**, NTHU, QUT English-to-Japanese ***LMAP:*** **OKSAT**, UKP, KMI ***Precision-at-5:*** **OKSAT**, KMI, UKP English-to-Korean ***LMAP:*** **KSLP**, OKSAT, KMI ***Precision-at-5:*** **KSLP**, OKSAT, KMI	English-to-Chinese ***LMAP:*** **KECIR**, QUT, KMI ***Precision-at-5:*** **OKSAT**, NTHU, QUT English-to-Japanese ***LMAP:*** **QUT**, UKP, OKSAT ***Precision-at-5:*** **OKSAT**, RDLL, UKP English-to-Korean ***LMAP:*** **KSLP**, KMI, OKSAT ***Precision-at-5:*** **KSLP,** KMI, OKSAT

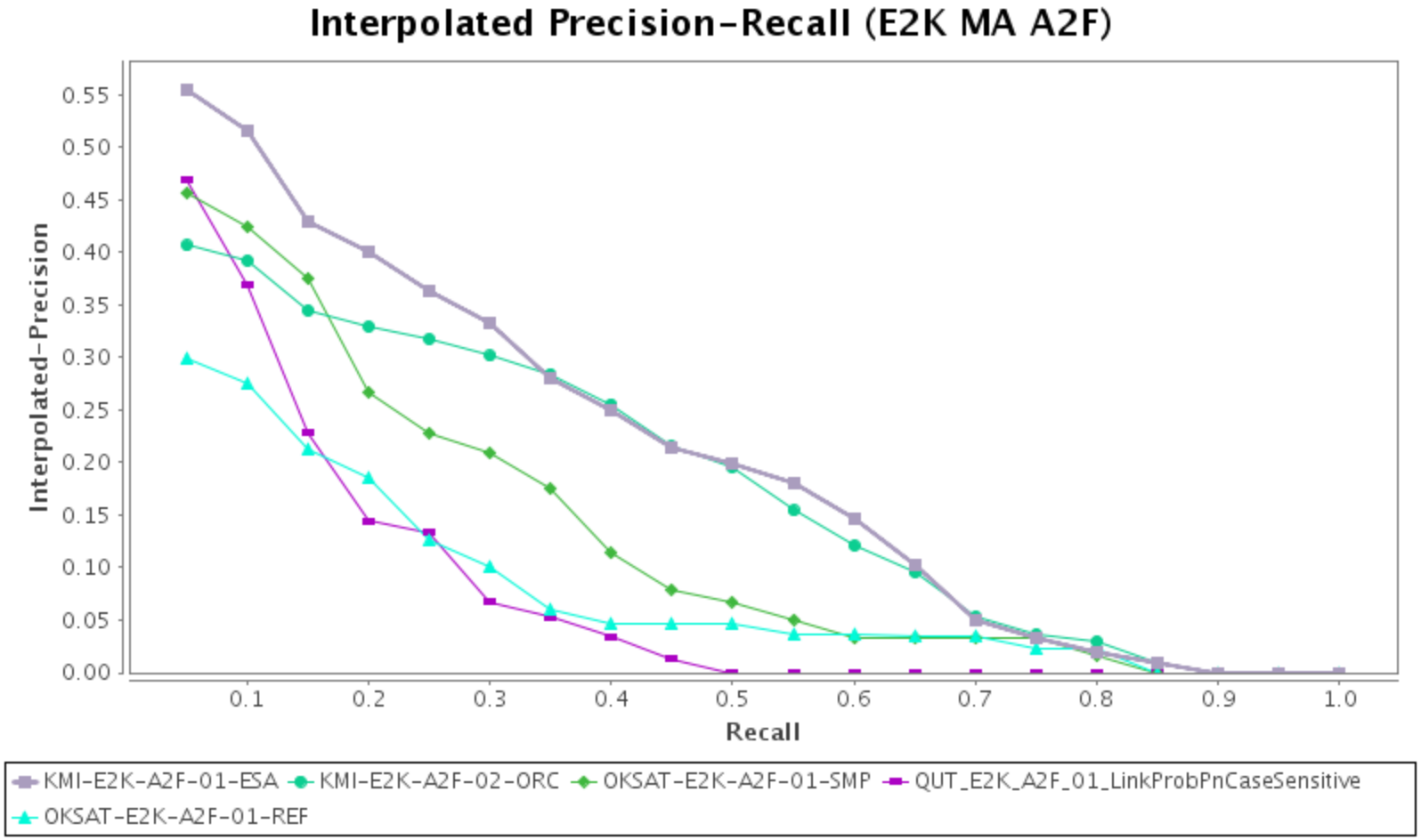

Interpolated Precision-Recall (E2C GT F2F)
Interpolated-Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
KMI-E2C-A2F-02-ORC
OKSAT-E2C-A2F-01-SMP
KMI-E2C-A2F-01-ESA
OKSAT-E2C-A2F-01-REF
QUT_E2C_A2F_01_LinkProbPnCaseSensitive
DCU-E2C-A2F-002-NWE
DCU-E2C-A2F-01-NW
III_E2C_A2F_01_PNM

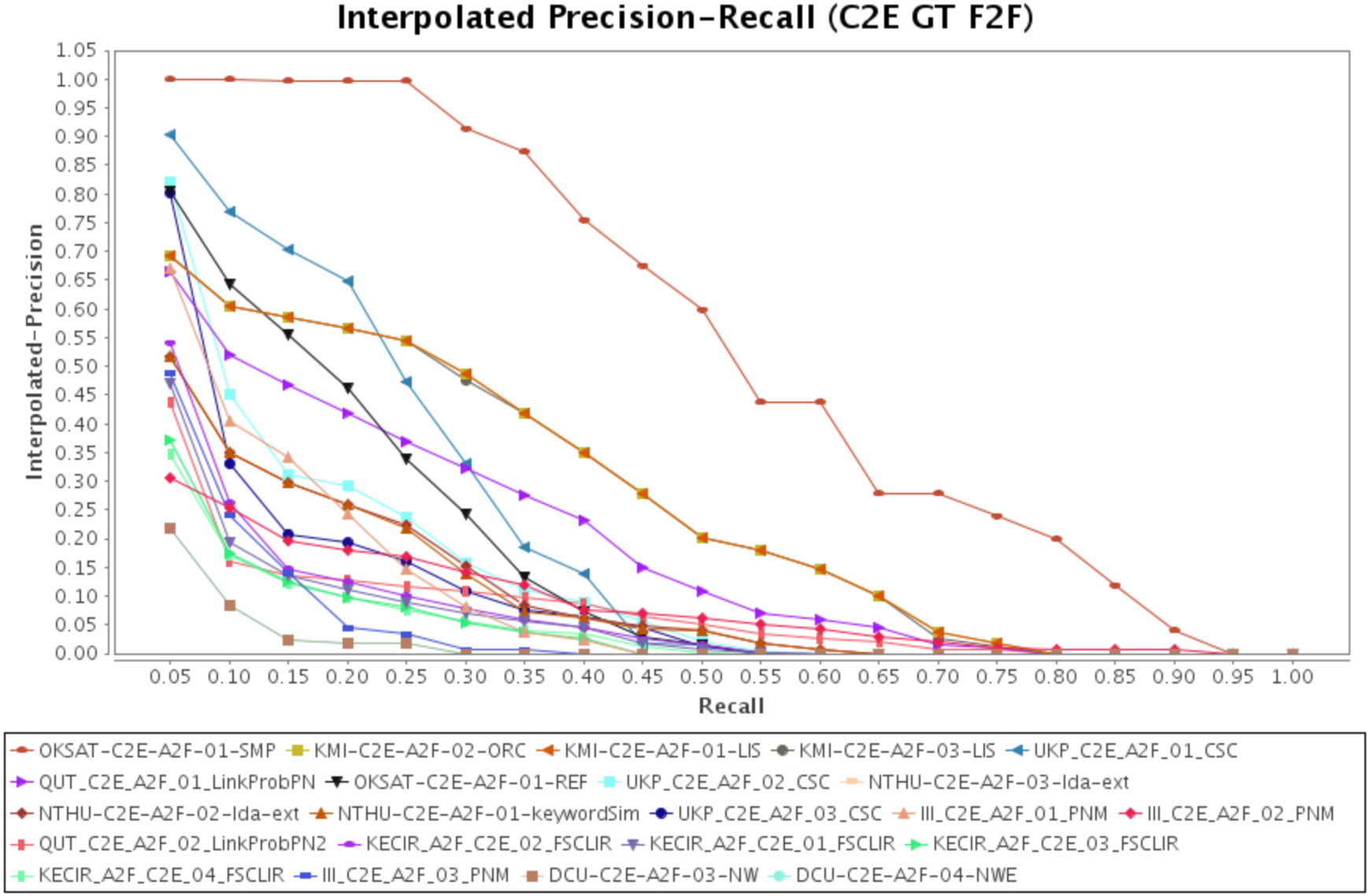

Interpolated Precision-Recall (E2C MA A2F)
Interpolated-Precision
0.375
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
QUT_E2C_A2F_01_LinkProbPnCaseSensitive
KMI-E2C-A2F-01-ESA
KMI-E2C-A2F-02-ORC
OKSAT-E2C-A2F-01-REF
OKSAT-E2C-A2F-01-SMP
DCU-E2C-A2F-01-NW
DCU-E2C-A2F-002-NWE
III_E2C_A2F_01_PNM

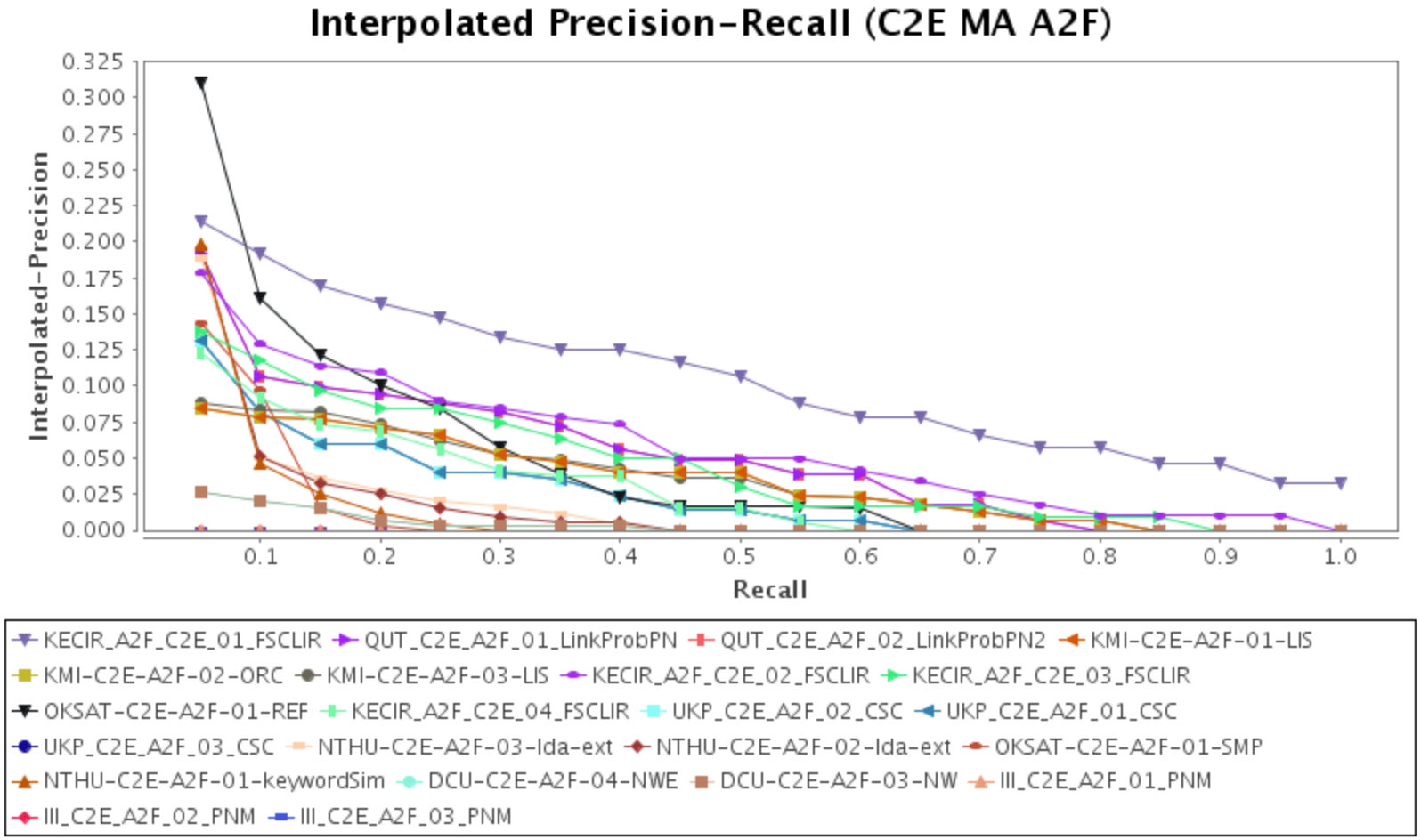

Interpolated Precision-Recall (E2J GT F2F)
Interpolated-Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
OKSAT-E2J-A2F-01-SMP
KMI-E2J-A2F-02-ORC
KMI-E2J-A2F-01-ESA
OKSAT-E2J-A2F-01-REF
QUT_E2J_A2F_01_LinkProbPnCaseSensitive

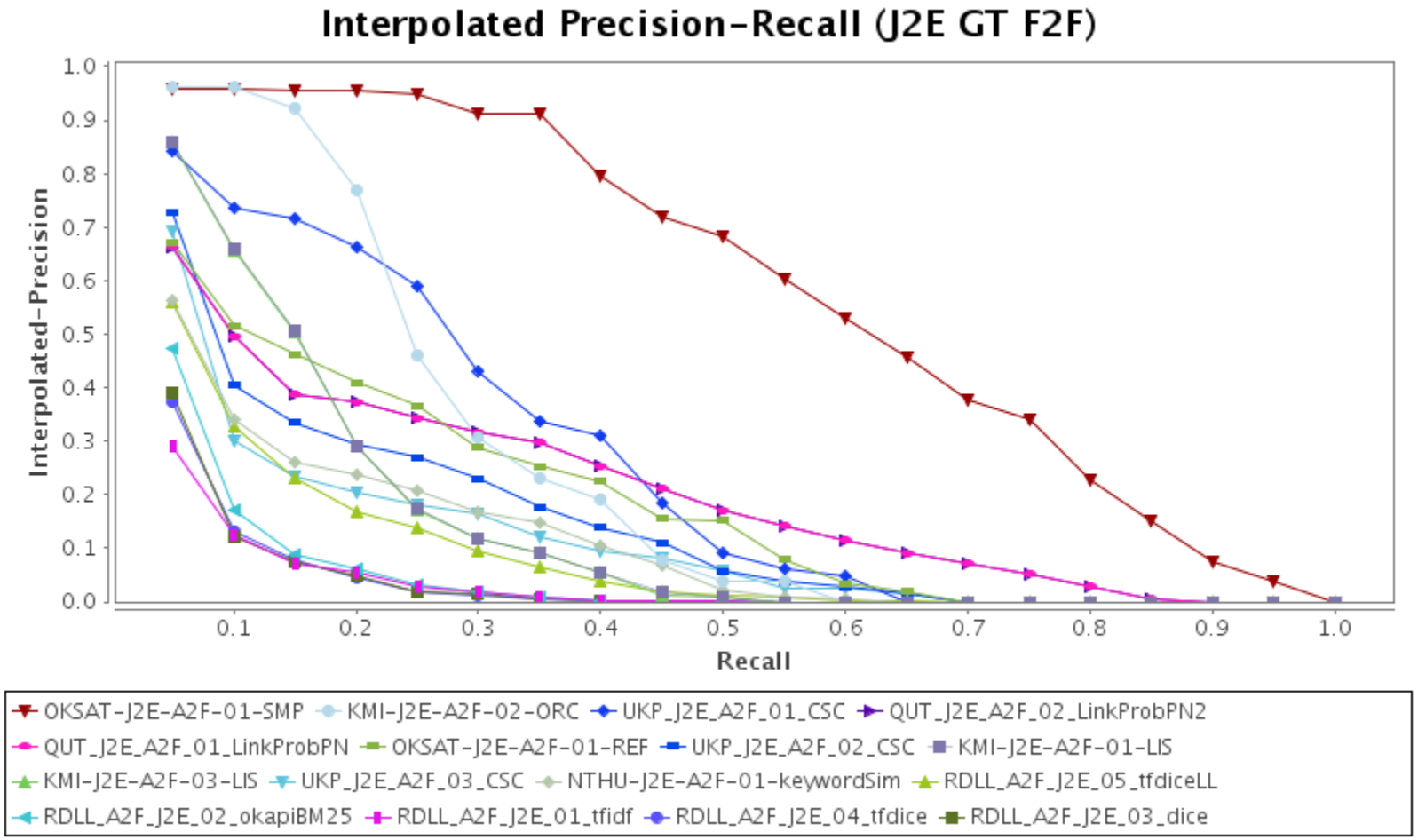

Interpolated Precision-Recall (E2J MA A2F)
Interpolated-Precision
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
KMI-E2J-A2F-02-ORC
KMI-E2J-A2F-01-ESA
OKSAT-E2J-A2F-01-SMP
QUT_E2J_A2F_01_LinkProbPnCaseSensitive
OKSAT-E2J-A2F-01-REF

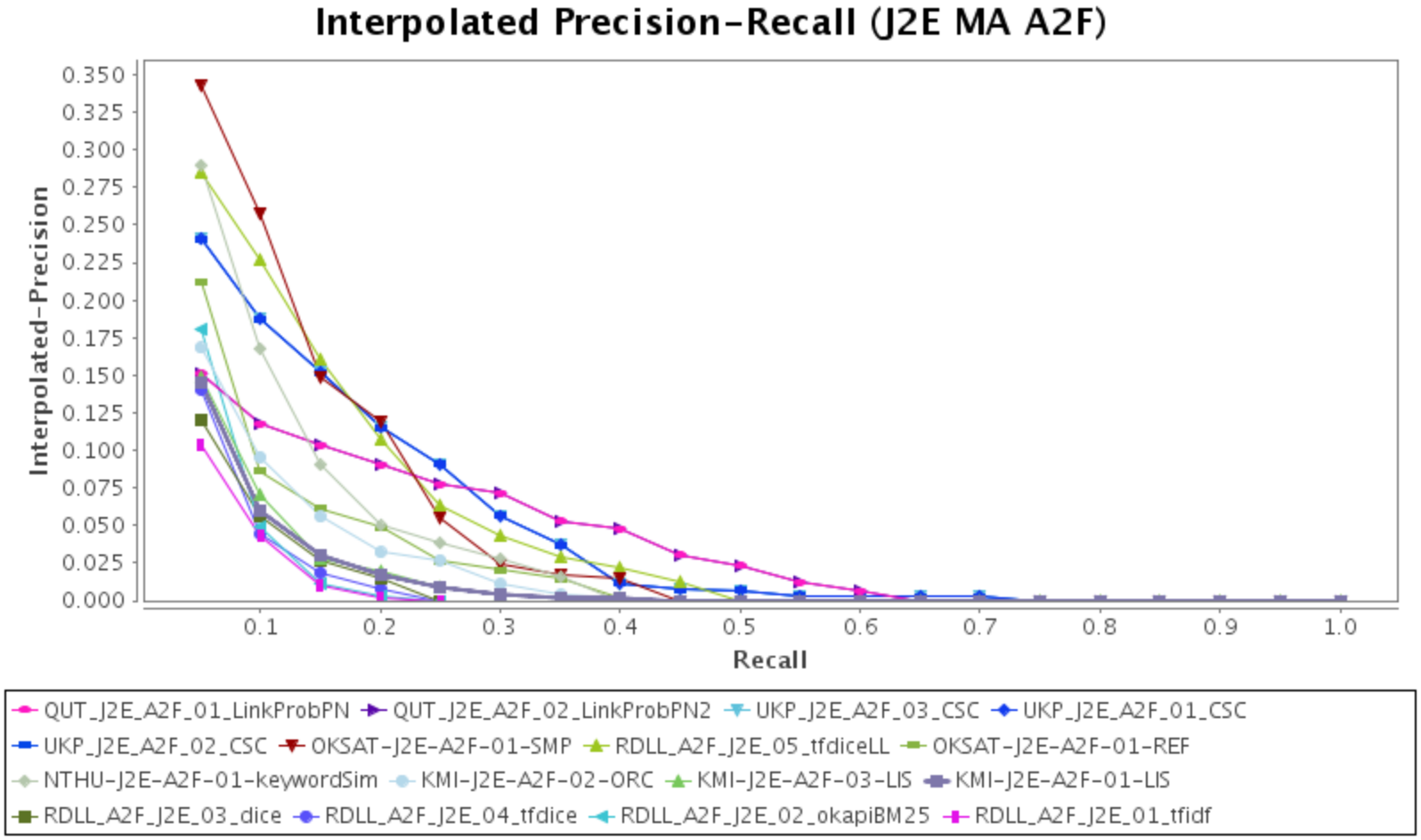

Interpolated Precision-Recall (E2K GT F2F)
Interpolated-Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
OKSAT-E2K-A2F-01-SMP
KMI-E2K-A2F-02-ORC
KMI-E2K-A2F-01-ESA
OKSAT-E2K-A2F-01-REF
QUT_E2K_A2F_01_LinkProbPnCaseSensitive

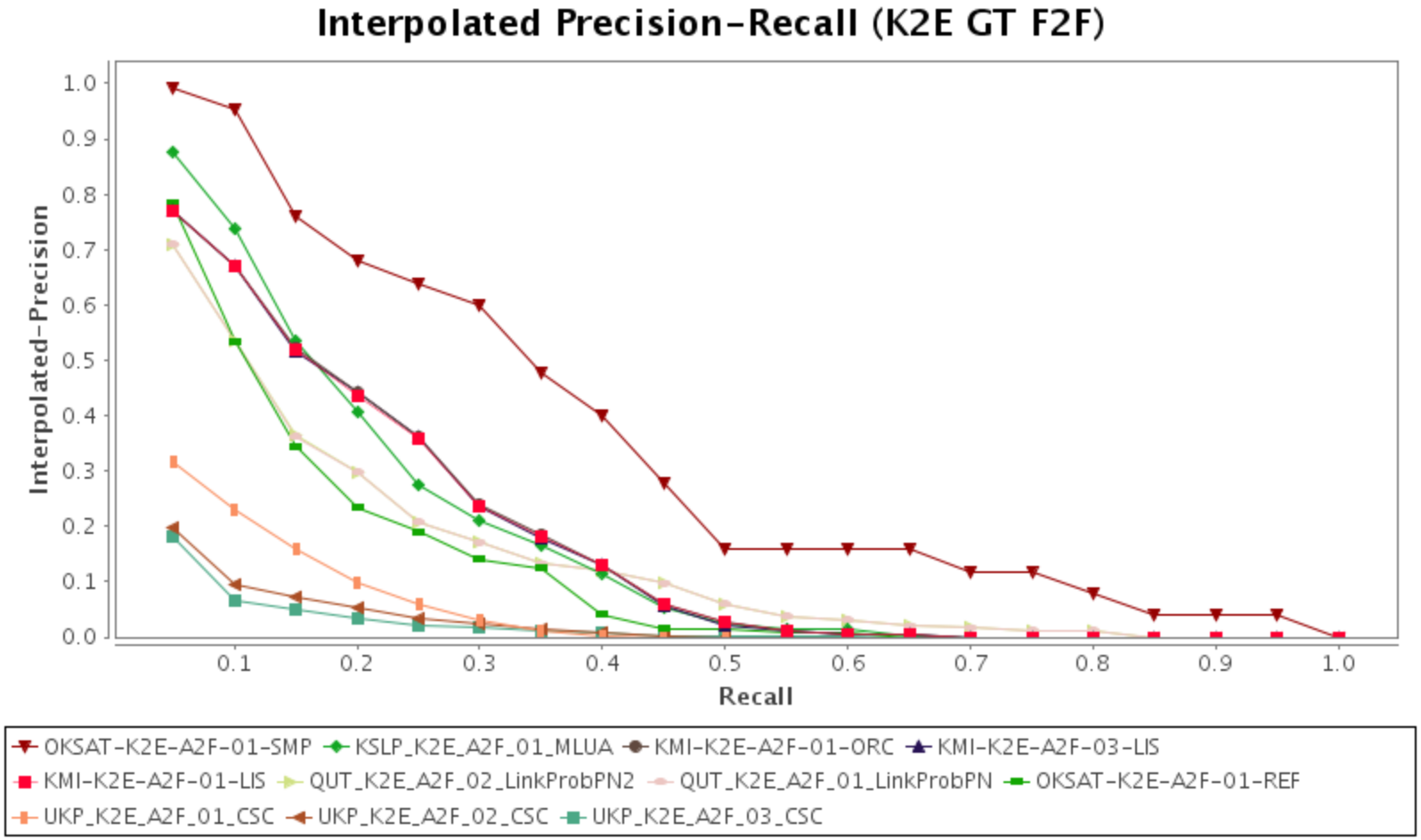

Interpolated Precision-Recall (E2K MA A2F)
Interpolated-Precision
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
KMI-E2K-A2F-01-ESA
KMI-E2K-A2F-02-ORC
OKSAT-E2K-A2F-01-SMP
QUT_E2K_A2F_01_LinkProbPnCaseSensitive
OKSAT-E2K-A2F-01-REF

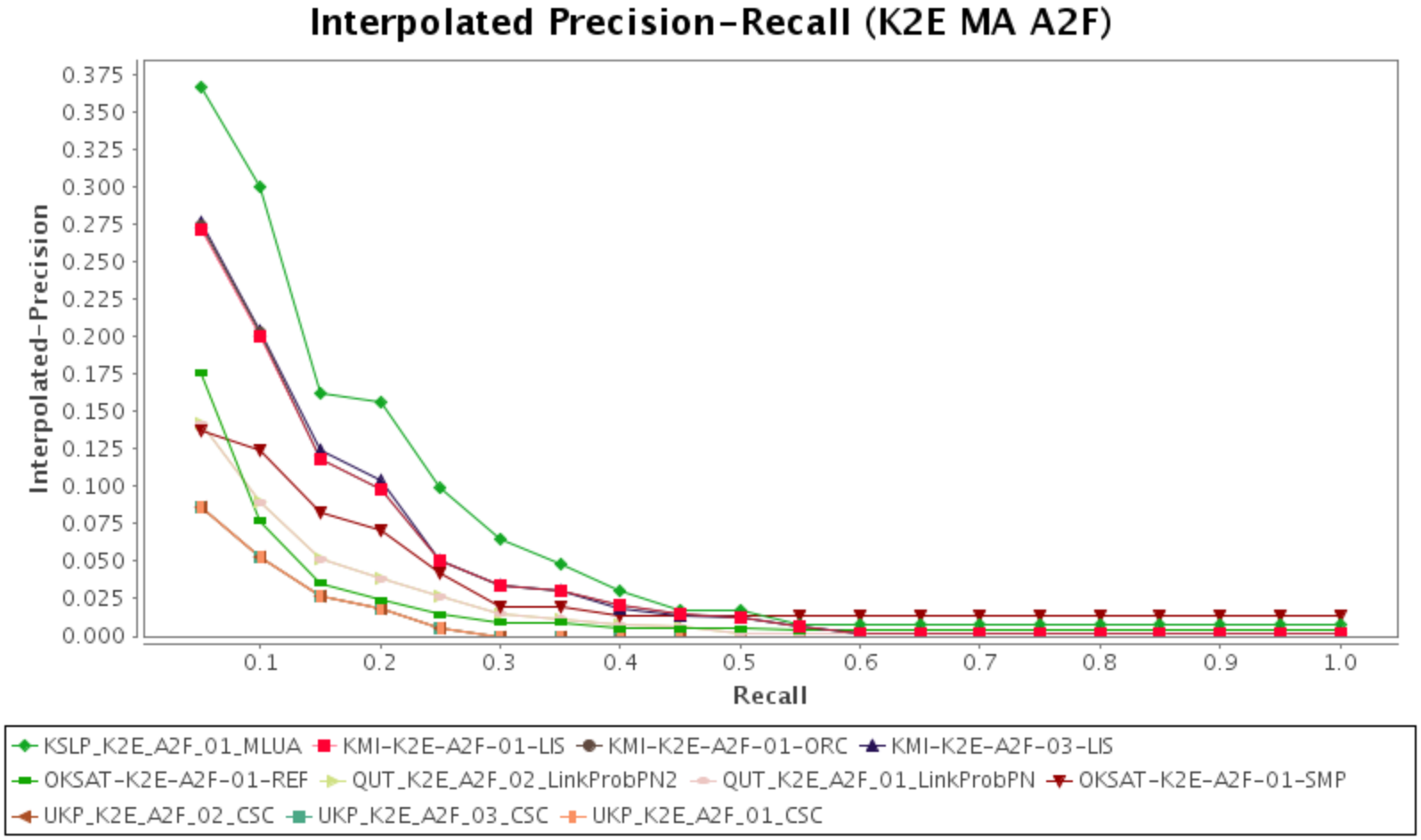

Interpolated Precision-Recall (C2E GT F2F)
Interpolated-Precision
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Recall
OKSAT-C2E-A2F-01-SMP
KMI-C2E-A2F-02-ORC
KMI-C2E-A2F-01-LIS
KMI-C2E-A2F-03-LIS
UKP_C2E_A2F_01_CSC
QUT_C2E_A2F_01_LinkProbPN
OKSAT-C2E-A2F-01-REF
UKP_C2E_A2F_02_CSC
NTHU-C2E-A2F-03-lda-ext
NTHU-C2E-A2F-02-lda-ext
NTHU-C2E-A2F-01-keywordSim
UKP_C2E_A2F_03_CSC
III_C2E_A2F_01_PNM
III_C2E_A2F_02_PNM
QUT_C2E_A2F_02_LinkProbPN2
KECIR_A2F_C2E_02_FSCLIR
KECIR_A2F_C2E_01_FSCLIR
KECIR_A2F_C2E_03_FSCLIR
KECIR_A2F_C2E_04_FSCLIR
III_C2E_A2F_03_PNM
DCU-C2E-A2F-03-NW
DCU-C2E-A2F-04-NWE


Interpolated Precision-Recall (C2E MA A2F)
Interpolated-Precision
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
KECIR_A2F_C2E_01_FSCLIR
QUT_C2E_A2F_01_LinkProbPN
QUT_C2E_A2F_02_LinkProbPN2
KMI-C2E-A2F-01-LIS
KMI-C2E-A2F-02-ORC
KMI-C2E-A2F-03-LIS
KECIR_A2F_C2E_02_FSCLIR
KECIR_A2F_C2E_03_FSCLIR
OKSAT-C2E-A2F-01-REF
KECIR_A2F_C2E_04_FSCLIR
UKP_C2E_A2F_02_CSC
UKP_C2E_A2F_01_CSC
UKP_C2E_A2F_03_CSC
NTHU-C2E-A2F-03-lda-ext
NTHU-C2E-A2F-02-lda-ext
OKSAT-C2E-A2F-01-SMP
NTHU-C2E-A2F-01-keywordSim
DCU-C2E-A2F-04-NWE
DCU-C2E-A2F-03-NW
III_C2E_A2F_01_PNM
III_C2E_A2F_02_PNM
III_C2E_A2F_03_PNM


Interpolated Precision-Recall (J2E GT F2F)
Interpolated-Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
OKSAT-J2E-A2F-01-SMP
KMI-J2E-A2F-02-ORC
UKP_J2E_A2F_01_CSC
QUT_J2E_A2F_02_LinkProbPN2
QUT_J2E_A2F_01_LinkProbPN
OKSAT-J2E-A2F-01-REF
UKP_J2E_A2F_02_CSC
KMI-J2E-A2F-01-LIS
KMI-J2E-A2F-03-LIS
UKP_J2E_A2F_03_CSC
NTHU-J2E-A2F-01-keywordSim
RDLL_A2F_J2E_05_tfdiceLL
RDLL_A2F_J2E_02_okapiBM25
RDLL_A2F_J2E_01_tfidf
RDLL_A2F_J2E_04_tfdice
RDLL_A2F_J2E_03_dice


Interpolated Precision-Recall (J2E MA A2F)
Interpolated-Precision
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
QUT_J2E_A2F_01_LinkProbPN
QUT_J2E_A2F_02_LinkProbPN2
UKP_J2E_A2F_03_CSC
UKP_J2E_A2F_01_CSC
UKP_J2E_A2F_02_CSC
OKSAT-J2E-A2F-01-SMP
RDLL_A2F_J2E_05_tfdiceLL
OKSAT-J2E-A2F-01-REF
NTHU-J2E-A2F-01-keywordSim
KMI-J2E-A2F-02-ORC
KMI-J2E-A2F-03-LIS
KMI-J2E-A2F-01-LIS
RDLL_A2F_J2E_03_dice
RDLL_A2F_J2E_04_tfdice
RDLL_A2F_J2E_02_okapiBM25
RDLL_A2F_J2E_01_tfidf


Interpolated Precision-Recall (K2E GT F2F)
Interpolated-Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
OKSAT-K2E-A2F-01-SMP
KSLP_K2E_A2F_01_MLUA
KMI-K2E-A2F-01-ORC
KMI-K2E-A2F-03-LIS
KMI-K2E-A2F-01-LIS
QUT_K2E_A2F_02_LinkProbPN2
QUT_K2E_A2F_01_LinkProbPN
OKSAT-K2E-A2F-01-REF
UKP_K2E_A2F_01_CSC
UKP_K2E_A2F_02_CSC
UKP_K2E_A2F_03_CSC


Interpolated Precision-Recall (K2E MA A2F)
Interpolated-Precision
0.375
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
KSLP_K2E_A2F_01_MLUA
KMI-K2E-A2F-01-LIS
KMI-K2E-A2F-01-ORC
KMI-K2E-A2F-03-LIS
OKSAT-K2E-A2F-01-REF
QUT_K2E_A2F_02_LinkProbPN2
QUT_K2E_A2F_01_LinkProbPN
OKSAT-K2E-A2F-01-SMP
UKP_K2E_A2F_02_CSC
UKP_K2E_A2F_03_CSC
UKP_K2E_A2F_01_CSC

Conclusions and Future Work

Answers of questions (1)

- Will natural language processing really help?
  - Not all teams used text segmentation for anchor identification
  - Team NTHU used a CKIP from Academia Sinica for Chinese segmentation
  - Team KECIR used FMM for Chinese segmentation
  - Team KSLP broke Korea text at '*eojeol*'

Segmentation seems helping, as team KECIR and KSLP achieved good A2F evaluation scores with manual assessment results in the C2E and K2E tasks separately.

Answers of questions (2)

- Will a unified linking method work on all kinds of cross-lingual link discovery with different link direction?

Mostly yes.

The top performer teams include KMI, OSTAT who employed a unified cross-lingual linking method achieved very good results in different language subtasks even with different link directions.

Conclusions

- Many good approaches were seen in the CJK to English cross-lingual link discovery tasks.
- The evaluation methods distinguish the effective and less effective CLLD algorithms.
- There are still lots of work needs to be done in the future.

Future works

- Personlised CLLD

  Just like other IR tasks, general approaches can't satisfy different needs.
- CLLD for other knowledge bases
- Patent <–> Wikipedia CLLD

  e.g. (lens.org <->Wikipedia)
- Patent CLLD
- …