KECIR at NTCIR-10 Cross-Lingual Link Discovery Task

Jianxi Zheng, Yu Bai, Cheng Guo, Dongfeng Cai
KERC, Shenyang Aerospace University, Shenyang, China
zhengjxkercir@163.com, baiyu@sau.edu.cn

ABSTRACT

This paper presents the methods of KECIR at NTCIR-10 Cross-Lingual Link Discovery Task. Two architectures of systems are designed, both of which consist of three common modules such as anchor detection, anchor translation and link discovery. In KECIR_A2F_C2E_03_FSCLIR and KECIR_A2F_C2E_04_FSCLIR, monolingual link discovery module is considered. In order to detect anchor, feature selection method is used. In the processing of anchor translation, we use a method combined with existing cross language link and Google translation web service. For discovering link, both title and paragraph matching methods are used to retrieve the relevant link corresponding to each anchor. Four runs were submitted, and in the A2F evaluation with Manual Assessment results, the KECIR_A2F_C2E_01_FSCLIR achieved the highest score of LMAP and R-Prec in Chinese to English task. The experiment shows that CrossLink based on the first architecture of system can retrieve higher precision links for an anchor than the second one, and anchors with noisiness will result in lower values of metrics in F2F evaluation.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – text analysis.
I.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – linguistic processing.

General Terms

Experimentation.

Keywords

Cross-Lingual Link Discovery, anchor detection, A2F evaluation, title and paragraph matching.

Team Name

KECIR.

Subtasks

Chinese to English CrossLink.

1. INTRODUCTION

The goal of Chinese-to-English CrossLink task [1] is aiming to automatically recognize anchors in the context of a Chinese test topic, and establish links to English Wikipedia Collections as soon as possible.
Wikipedia is the largest multi-lingual encyclopedia in the world. However, majority of links in Wikipedia articles are monolingual links and there are only a few cross-lingual links across documents in different languages. This leads serious difficulties to users who try to read articles written in other languages to extend the understanding for a word without the monolingual explanation. Therefore, the technology of CrossLink can break language barrier and realize knowledge discovery across documents in different languages.
Feature selection method and cross-language information retrieval (CLIR) approach based on anchor translation are used in our work. We suppose that an anchor is a feature for each test topic, and the anchor detection should be regarded as a problem of feature selection. For identifying anchor, a Chinese segmentation technology based on existing structure information of Wikipedia is developed to split consecutive string in Chinese articles into single words. After detecting anchor, anchor translation based CLIR method can be explored to retrieve cross-lingual links corresponding to an anchor. Overall, we divide this subtask into feature selection and Cross-Language Information Retrieval.
The remainder of the paper is organized as follows: in Section 2, related work about CLLD will be introduced. The architectures of systems and all modules of which are given in from Section 3 to Section 6, respectively. The experiment in our work is described in Section 7. Section 8 will conclude our work and give our plan about CLLD in future.

2. RELATED WORK

Previous works on Cross-Lingual Link Discovery mainly focused on Links from English to other Language documents, which include the followings. Sorg and Cimiano [2] proposed a classification-based approach to obtain new cross-language links between English and German Wikipedia articles. Smet and Moens [3] used interlingual topic model to find English to Dutch Linking of News Stories on the Web. In the NTCIR-9, English to CJK (Chinese/Japanese/Korean) tasks were first proposed [4] and attracted some teams like [5-15] interested with CrossLink. Especially, considering at which stage to carry out the cross-lingual matching, Kim and Gurevych [5] summarized three different approaches to CrossLink such as document translation based CLLD, anchor translation based CLLD, and cross-lingual document similarity based CLLD. In this paper, anchor translation based CLLD and cross-lingual document similarity based CLLD are utilized in our systems. However, in 2012, Tang

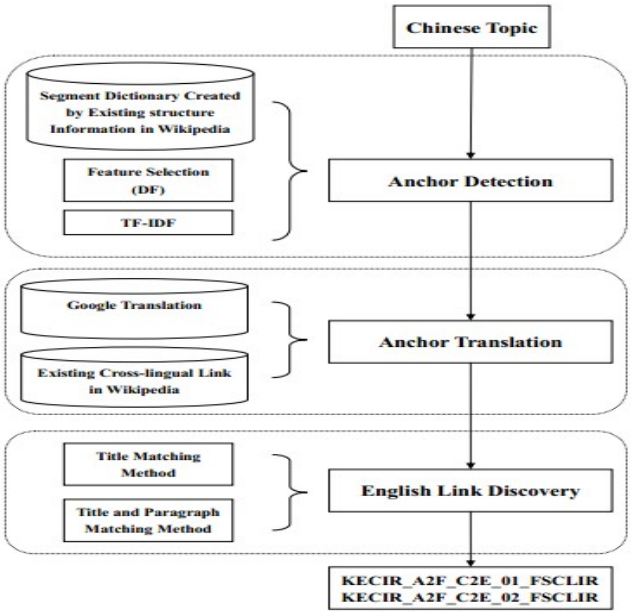

Figure 1: Architecture of System I

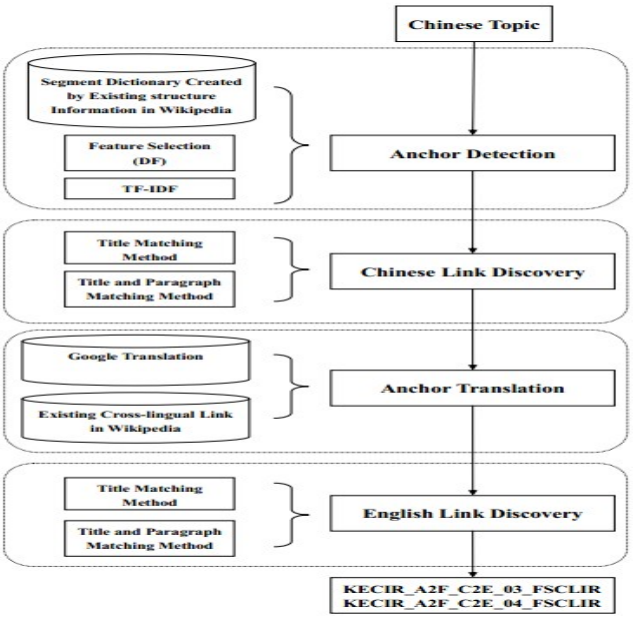

Figure 2: Architecture of System II

[16] presented relevant approach to automatic Chinese to English Cross-lingual link discovery, in which they studied the effects of Chinese segmentation and Chinese to English translation on the hyperlink recommendation.

3. SYSTEM ARCHITECTURE

Two architectures of systems are designed, both of which consist of three common modules such as anchor detection, anchor translation and link discovery. Figure 1 and 2 describe the details of them. In KECIR_A2F_C2E_03_FSCLIR and KECIR_A2F_C2E_04_FSCLIR, monolingual link discovery module is considered. In order to detect anchor, feature selection method is employed. And then, we use a method combined with existing cross language link in Chinese training Wikipedia articles and Google translation web service in the processing of anchor translation. For discovering link, both title and paragraph matching methods are exploited to retrieve at most 5 relevant links corresponding to each translated anchor.

4. ANCHOR DETECTION

4.1 Test Topics and Training Corpus Preprocessing

According to the first rule required in NTCIR-10 websites that only main text of test topics between <bdy> and reference or external links section can be linked, we first extract the main parts starting after <bdy> tag and ending before three special tags such as <st>References</st>, <st>External Links</st> and <st>See Also</st> from each of twenty-five Chinese test topics. It is worthy of noticing that the special ending tags in English Wikipedia articles maybe corresponds to multiple tags in Chinese ones, which are listed in Table 1.

4.2 Word Segmentation with Existing Structure of Wikipedia

As we all known that there are no word boundaries in Chinese topics, so a proper Chinese segment technology for Wikipedia articles is needed to solve the problem. Here we choose the Forward Maximum Matching (FMM) [17] algorithm because it is a simple and effective method which only needs a dictionary. So how to create a dictionary becomes a first challenge task for segmenting word. It is difficult to construct a dictionary for satisfying with the segment need in a Chinese Wikipedia article not only because many Chinese Wikipedia articles are written in a form of mix of simplified, traditional Chinese writing, but also because some words in Wikipedia are out of vocabulary terms. In order to alleviate the problems, we create a Wikipedia-based dictionary containing existing structure information in Wikipedia articles. Through observing some Chinese Wikipedia articles, we find two important elements with tag such as texts between tag <title> and </title>, and ones between and . Take the Chinese topic “马王堆汉墓” as an example, tags <title>马王堆汉墓</title> and 马鞍堆 are important structure information in a manually edited Wikipedia articles, so “马王堆汉墓” and “马鞍堆” are important terms which will be stored in the dictionary. We extract all such structure information from Chinese Wikipedia topics and training corpus, filter some noises like punctuations, and finally create a dictionary containing 349,428 terms in total. Then FMM algorithm based on the Wikipedia-based dictionary is used to split consecutive string in Chinese text into separate words.

4.3 Feature Selection Approach to Anchor Detection

Feature Selection methods including DF (Document Frequency), MI (Mutual Information), CHI, and IG (Information Gain), etc have been discussed on text classification in recent years [18]. In this short paper, we choose DF method to identify the anchor for each topic. After breaking Chinese

Table 1: multiple tags in Chinese articles corresponding to three special tags in English ones

English	Chinese
<st>References</st>	<st>参考资料</st>, <st>参考</st>, <st>参考及来源</st>, <st>引用</st>, <st>参考文献</st>.
<st>External Links</st>	<st>外部链接</st>, <st>外部连接</st>, <st>外部连结</st>.
<st>See Also</st>	<st>参看</st>, <st>参见</st>.

text into separate words, the Document Frequency (DF) of each separate word in both Chinese test topic and training corpus can be obtained and then we use the TF*IDF schema [19] to weight the importance of separate words in each topic. After filtering chronological items such as number, or year links required at NTCIR-10 websites, the remainder of words are regarded as candidate anchors in each topic. There are at most 250 anchors for each topic setting up in processing of detecting anchor.

5. ANCHOR TRANSLATION

This section introduces our method of translating anchor. We first extract the existing cross-language link from Chinese document collection as translation dictionary. And if there is no translation for an anchor in the dictionary, Google translation web service [20] is used.

6. LINK DISCOVERY

In this section, an Information Retrieval Platform including index creation and link discovery modules is developed to retrieve relevant links for the corresponding anchor per topic.

For creating index, we first split the English raw corpora into several short paragraph texts by tag <sec> using regular expressions. Each of texts is represented as the predefined XML styles shown in Table 2, which consists of four elements such as title, id, category and body. The first element represents title of raw corpora. The second element is id of texts which is represented as id-k, where id equals with the number between tag <id> and </id> in raw corpora and k varies from 1 to the total number of tag <sec> in one. The third element represents categories of texts, which combines all categories of raw corpora using tag "|" as separate symbol. The final element is the body parts of short text, which is represented as several paragraphs of raw corpora.

Table 2: the XML Style for English document preprocessing

<title>title of raw topic</title>
<id>$id-k$</id>
<category>$category_1$|$category_2$|..|$category_n$</category>
<bdy>text of raw test topic between tag
$<bdy>$ (or $</sec>$) and $<sec>$</bdy>

For discovering link, we exploit two methods to retrieve at most 5 relevant English documents for a translated anchor. The first method is a title matching approach which aims at finding relevant titles of page for a translated anchor. In this method, Vector Space Model (VSM) API in Lucene [21] is used to retrieve top5-ranked English documents for a translated anchor. The second method is a title and paragraph matching method which refers to retrieve title and paragraph at the same time for a translated anchor, the goal of which is to explore which page and its paragraph are linked to the corresponding anchor. This method mainly contains three steps:

Step1: VSM is used to retrieve top100-ranked English paragraphs for a translated anchor.

Step2: for an English document, the relevance with corresponding translated anchor is calculated by summing up weights of all paragraphs it contains.

Step3: sort the relevance between the English document and its corresponding anchor by descending, at most 5 relevant English documents for an anchor are returned.

Finally, our four runs are generated by using combination of methods stated in above sections, and the details of which are given in Table 3.

7. EXPERIMENT

7.1 Evaluation Results

For C-2-E CrossLink subtask, four runs were finally submitted and the performances of which were evaluated using LMAP metric (Link Mean Average Precision), R-Prec metric, and Precision-@-N (N is 5,10,20,30,50,250) in both File-to-File (F2F) and Anchor-to-File (A2F) with Wikipedia Ground Truth and Manual Assessment respectively, and all of them are described in Table 4. Besides, the Interpolated Precision-Recall curves of four runs in three different evaluation scenarios are also given in Figure 3, 4, 5.

7.2 Analysis for the performance of submissions

In order to facilitate the description, we respectively call four submissions including KECIR_A2F_C2E_01_FSCLIR, KECIR_A2F_C2E_02_FSCLIR, KECIR_A2F_C2E_03_FSCLIR, and KECIR_A2F_C2E_04_FSCLIR as Run 1, Run 2, Run 3 and Run 4. When runs are measured in A2F level with Manual Assessment, Run 1 gets the top score of LMAP and R-Prec outperformed other three runs. It is out of our expectation that the performance of the second run with content information is lower than one of the first run. When evaluated with Wikipedia Ground Truth in F2F level, Run 2 achieves higher performance than Run 1, which shows that content information helps discover articles of same topic and find more existing cross language links in Wikipedia. In F2F evaluation with Manual Assessment, Run 3 performs better than the remainder of runs in metric of Precision-at-5 while it ranks third in four runs in scores of LMAP and R-prec. From figure 3 to figure 5, we conclude that the former two runs across all evaluation perform better than the rest of runs, which proves the performance of system I is higher than one of system II.

7.3 Performance comparison with other teams

In A2F evaluation with Manual Assessment, our run KECIR_A2F_C2E_01_FSCLIR achieves the top scores of LMAP and R-prec outperformed other teams, and ranks the eighth at official results in terms of Precision-at-N. It demonstrates that our method can effectively identify anchor and recommend relevant links for it in Chinese topic.However, comparison with the best official results, the performances of four runs are lower in both of in F2F evaluation with Wikipedia Ground Truth and Manual Assessment results.

8. CONCLUSIONS AND FUTURE WORK

This paper describes the methods of KECIR at NTCIR-10 CrossLink-2 Task. The KECIR_A2F_C2E_01_FSCLIR achieved the best score of LMAP and R-Prec with Manual Assessment in A-2-F evaluation in Chinese to English task. This demonstrates our method can effectively recommend relevant links for each anchor per test topic. However, our four runs don't perform so well in both of in F2F evaluation with Wikipedia Ground Truth and Manual Assessment results. It proves that our method needs to promote the ability of finding articles of same topic from multilingual Wikipedia articles. The experiment shows that CrossLink based on the first architecture of system can retrieve higher precision links for an anchor than the second one, and anchors with noisiness will result in lower values of metrics in F2F evaluation. We plan to build a united framework for Cross-lingual Link between Chinese articles and English ones in future. Other feature selection approaches to anchor detection and anchor translation based on Web will be explored in the united framework of CrossLink.

9. ACKNOWLEDGEMENTS

Thanks to the organizers of NTCIR-10 CrossLink-2 track. And, our work was supported by NSFC 61073123, The National Key Technology R&D Program 2012BAH14F00, 973 Program of China 2010CB530401.

10. REFERENCES

[1] L.-X. Tang, I.-S. Kang, F. Kimura, Y.-H. Lee, A. Trotman, S. Geva and Y. Xu, "Overview of the NTCIR-10 Cross-Lingual Link Discovery Task," In proceedings of NTCIR-10, Tokyo, Japan, 2013, pp 1-36.

[2] P. Sorg and P. Cimiano, "Enriching the Crosslingual Link Structure of Wikipedia - A Classification-Based Approach," In proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, 2008, pp 49-54.

[3] W. D. Smet and M. F. Moens, "Cross-Language Linking of News Stories on the Web Using Interlingual Topic Modelling," In proceedings of SWSM' 09. Hong Kong, China, 2009, pp 1-8.

[4] L.-X. Tang, S. Geva, A. Trotman, Y. Xu, and K. Y. Itakura, "Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery," in Proceedings of NTCIR-9, Tokyo, Japan, 2011, pp. 437-463.

[5] J. Kim and I. Gurevych, "UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 487-494.

[6] P. Knoth, L. Zilka, and Z. Zdrahal, "KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 495-502.

[7] A. Fahrni, V. Nastase, and M. Strube, "HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 473-480.

[8] L.-X. Tang, D. Cavanagh, and A. Trotman, "Automated Cross-lingual Link Discovery in Wikipedia," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 512-529.

[9] P. H. Anh and T. Yukawa, "Using Concept base and Wikipedia for Cross-Lingual Link Discovery," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 464-468.

[10] C.-Y. Cheng, Y.-C. Wang and R. T.-H. Tsai, "IISR Crosslink Approach at NTCIR 9 CLLD Task," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 469-472.

[11] I.-S. Kang and R. Marigomen, "English-to-Korean Cross-linking of Wikipedia Articles at KSLP," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 481-483.

[12] M. F. Liu, L. Kang, and S. Yang, et al. "WUST EN-CS Crosslink System at NTCIR-9 CLLD Task," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 508-511.

[13] Y. F. Gao, H. J. Xu, and J. S. Zhang, et al. "Multi-filtering Method Based Cross-lingual Link Discovery," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 520-523.

[14] S.-J. Kang, "Cross-lingual Link Discovery by Using Link Probability and Bilingual Dictionary," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 484-486.

[15] Y. H. Lee, C. Y. Chuang, and C. C. Chen, "Discovering Links by Context Similarity and Translated Key Phrases for NTCIR9 CrossLink," In proceedings of NTCIR-9, Tokyo, Japan, 2011, pp 503-507.

[16] L.-X. Tang, A. Trotman, S. Geva, Y, Xu. "Cross-Lingual Knowledge Discovery: Chinese-to-English Article Linking in Wikipedia," In proceedings of the Eighth Asia Information Retrieval Societies Conference, 2012, pp 1-10.

[17] Y. Liu, Q. Tan, and K. X. Shen. "The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing (in Chinese)," Qing Hua University Press and Guang Xi Science and Technology Press, 1994, pp 36.

[18] S.-S Li, R. Xia, C.-Q. Zong, and C.-R. Huang, "A Framework of Feature Selection Methods for Text Categorization," In proceedings of the 47th Annual Meeting of the ACL and the AFNLP, 2009, pp 692-700.

[19] S. Robertson, "Understanding Inverse Document Frequency: On theoretical arguments for IDF," Journal of Documentation, 2004, pp 1-19.

[20] http://translate.google.cn

[21] http://lucene.apache.org

Table 3: the descriptive of four runs

Run ID	descriptive
KECIR_A2F_C2E_01_FSCLIR	FMM+TF-IDF+anchor translation+link discovery (title matching).
KECIR_A2F_C2E_02_FSCLIR	FMM+TF-IDF+anchor translation+link discovery (title and paragraph matching).
KECIR_A2F_C2E_03_FSCLIR	FMM+TF-IDF+Chinese link discovery(title matching) and anchor translation+link discovery(title matching).
KECIR_A2F_C2E_04_FSCLIR	FMM+TF-IDF+Chinese link discovery(title and paragraph matching) and anchor translation+link discovery(title and paragraph matching).

Table 4: the performance of four runs

Run ID	LMAP	R-Prec	P5	P10	P20	P30	P50	P250
the performance of four runs in F2F evaluation with Wikipedia Ground-Truth								
KECIR_A2F_C2E_01_FSCLIR	0.046	0.105	0.136	0.128	0.138	0.132	0.110	0.055
KECIR_A2F_C2E_02_FSCLIR	**0.054**	**0.119**	**0.200**	**0.164**	**0.150**	**0.145**	**0.126**	**0.060**
KECIR_A2F_C2E_03_FSCLIR	0.036	0.097	0.080	0.108	0.106	0.107	0.102	0.052
KECIR_A2F_C2E_04_FSCLIR	0.036	0.105	0.080	0.112	0.110	0.116	0.110	0.054
the performance of four runs in F2F evaluation with Manual Assessment								
KECIR_A2F_C2E_01_FSCLIR	**0.044**	**0.081**	0.080	0.088	0.094	**0.100**	0.090	**0.084**
KECIR_A2F_C2E_02_FSCLIR	0.037	0.077	0.072	0.092	**0.096**	0.095	**0.091**	0.072
KECIR_A2F_C2E_03_FSCLIR	0.031	0.076	**0.096**	**0.100**	0.088	0.081	0.083	0.071
KECIR_A2F_C2E_04_FSCLIR	0.028	0.076	0.056	0.096	0.088	0.084	0.086	0.066
the performance of four runs in A2F evaluation with Manual Assessment								
KECIR_A2F_C2E_01_FSCLIR	**0.087**	**0.054**	**0.024**	**0.036**	**0.046**	**0.061**	**0.074**	**0.064**
KECIR_A2F_C2E_02_FSCLIR	0.050	0.039	**0.024**	**0.036**	0.042	0.047	0.055	0.047
KECIR_A2F_C2E_03_FSCLIR	0.044	0.032	0.016	0.020	0.024	0.036	0.046	0.040
KECIR_A2F_C2E_04_FSCLIR	0.029	0.025	0.016	0.020	0.028	0.037	0.038	0.032

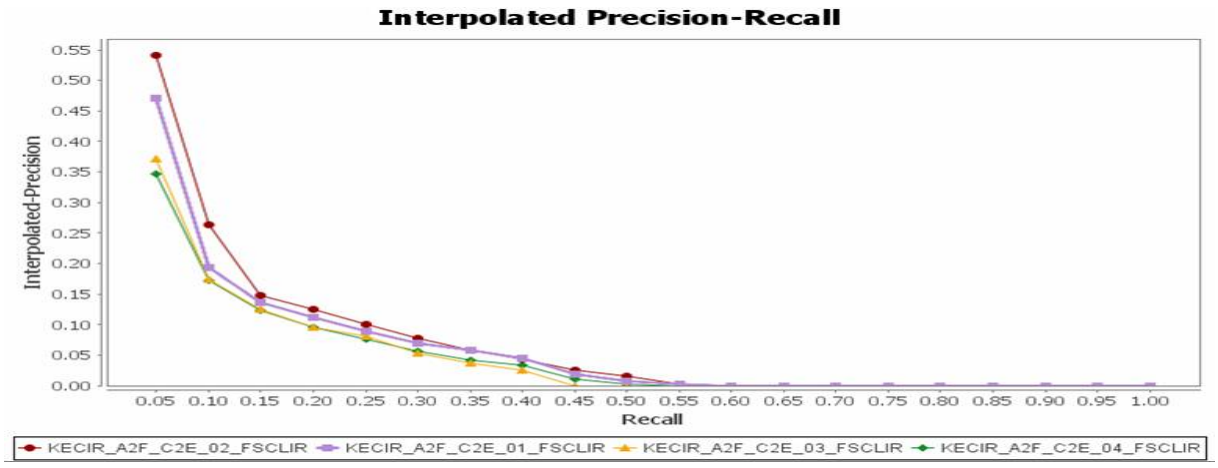

Figure 3: Interpolated Precision-Recall curves of four runs in F2F evaluation with Wikipedia Ground-Truth

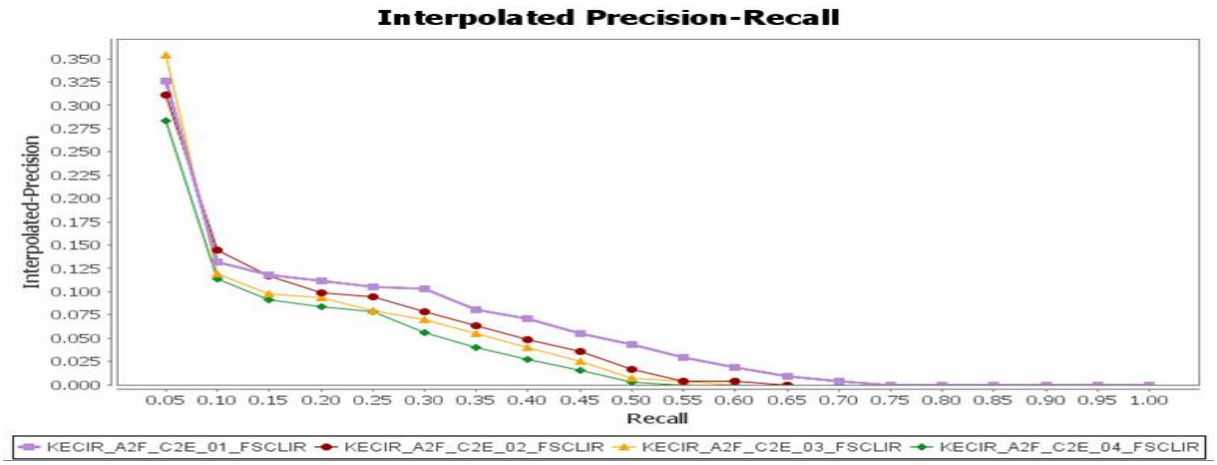

Figure 4: Interpolated Precision-Recall curves of four runs in F2F evaluation with Manual Assessment results

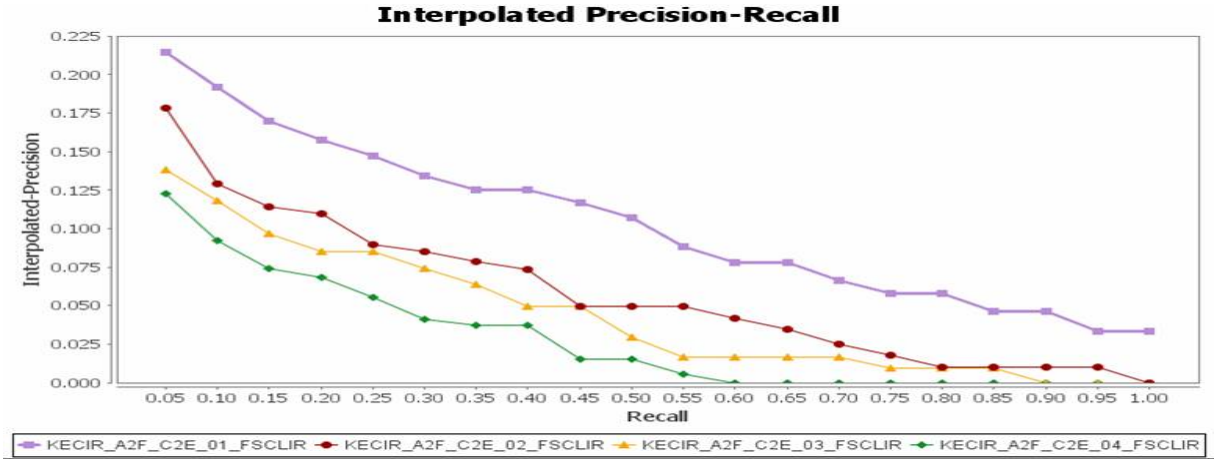

Figure 5: Interpolated Precision-Recall curves of four runs in A2F evaluation with Manual Assessment results