NTCIR-5 CLIR Experiments at Oki

Tetsuji Nakagawa
nakagawa378@oki.com
Corporate Research and Development Center
Oki Electric Industry Co., Ltd.
2−5−7 Honmachi, Chuo-ku, Osaka 541-0053, Japan

Abstract

We participated in the SLIR, BLIR(PLIR) and MLIR subtasks of the NTCIR-5 CLIR task. Our IR system uses language models for document scoring and query expansion, and can handle four languages; Chinese, Japanese, Korean and English. The system utilizes multiple language resources (bilingual dictionaries, parallel corpora and machine translation systems). We attempted to use some techniques including pivot language approach, query translation and document translation.

Keywords: *language models, multilingual information retrieval, pivot language, query translation, document translation.*

1 Introduction

We participated in the SLIR, BLIR(PLIR) and MLIR subtasks of the NTCIR-5 CLIR task. Our main goal is to develop a CLIR system which can handle as many languages as possible even with limited available resources for translation. The IR system we developed conducts document scoring and pseudo-relevance feedback based on language models, and uses several resources for cross-lingual IR; bilingual dictionaries, parallel corpora and machine translation systems. We attempt to use some techniques including the pivot language approach, query translation and document translation. Our system can handle four types of queries; queries written in Chinese, Japanese, Korean and English, and five types of documents; documents written in the four languages plus multilingual documents (CJKE). We submitted search results for all the 20 subtasks of these combinations ({C, J, K, E}-{C, J, K, E, CJKE} runs).

This paper is organized as follows: Section 2 describes our IR system. Section 3 discusses results of the formal runs and Section 4 concludes.

2 System Description

The system uses word-based indexing for Chinese, Japanese, Korean and English. Language models are used for document scoring, and the pseudo-relevance feedback is used for query expansion. In bilingual IR, the cross-lingual pseudo-relevance feedback method is used for query translation. In multilingual IR, each result of SLIR and BLIR for the same query is merged with a normalizing method. We explain these methods in the following subsections.

2.1 Keyword Extraction

Two major approaches are known for keyword extraction from queries and documents written in Chinese and Japanese — the n-gram-based approach and the word-based approach. The n-gram-based approach uses character n-grams for indexing, and needs no language-specific word segmenters. The word-based approach needs smaller size of indices, and a word is a suitable unit for linguistic processing than a character n-gram. We adopt the word-based approach. We developed a statistical Chinese word segmenter [8] and a statistical Korean morphological analyzer, and use them for Chinese and Korean keyword extraction respectively. The Japanese morphological analyzer ChaSen [7] is used for Japanese keyword extraction, and the Porter stemmer is used for English. All symbol characters are removed from indices. Functional words in Japanese indices identified by the morphological analyzer and stopwords (429 words) in English indices are also removed.

2.2 Document Scoring

We use the language models [11] for document scoring. Given a query q and a document d, the method uses the probability of d given q as the relevance between q and d. We use the following equation ($score_4$ model described in [4]) to calculate the retrieval status value of the document, $RSV(d)$ (see Appendix A for more details):

$$
\begin{aligned}
RSV(d) &= \log P(d|q), \\
&\simeq \log \sum_{t'} tf(t',d) + \\
&\sum_{t \in q \cap d} tf(t,q) \log \left\{ \frac{\lambda tf(t,d) \sum_{t'} df(t')}{(1-\lambda) df(t) \sum_{t'} tf(t',d)} + 1 \right\} + c. \quad (1)
\end{aligned}
$$

where $tf(x,y)$ is the frequency of the term x in the query or document y, $df(x)$ is the number of documents containing the term x, and λ is a smoothing parameter ranging from 0 to 1. c is a constant independent of d and ignored in the calculation of the RSV. The $RSV(d)$ can be calculated by using an inverted file, and we use Generic Engine for Transposable Association (GETA) [5] for this purpose. We set the value of the smoothing parameter λ to 0.25, which yields high performance on the training(NTCIR-4) data.

2.3 Query Expansion

We use the pseudo-relevance feedback (PRF) method to expand queries. Given a query, the method retrieves the top M documents $\mathbf{r} = \{r_1, \cdots, r_M\}$. Each term t in the documents are ranked by a certain term selection value $TSV(t)$ and the top N terms are added to the initial query. We use the ratio method [10] for the scoring. This method calculates the probability that $\mathbf{r}$ is generated given the term t based on language models and uses it as the term selection value $TSV(t)$ as follows:

$$
\begin{aligned}
TSV(t) &= \log P(\mathbf{r}|t), \\
&\simeq \log \prod_{d' \in \mathbf{r}} \frac{P(t|d')P(d')}{P(t)}, \\
&= \sum_{d' \in \mathbf{r}} \log \frac{P(t|d')}{P(t)} + c. \qquad (2)
\end{aligned}
$$

where c is a constant independent of t. We set the values of M and N both to 10, which performed well on the training data.

2.4 Cross-Lingual IR

We prepared the following language resources for translation:

Bilingual Dictionary

- *CEDICT* — Chinese-English dictionary (27,085 words) [3]
- *EDICT* — Japanese-English dictionary (110,872 words) [1]
- *engdic* — English-Korean dictionary (212,699 words) [9]

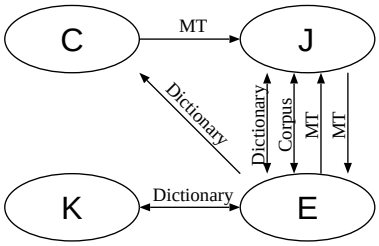

Figure 1. Translation Resources

Parallel Corpus

- *Japanese-English News Article Alignment Data* — Japanese and English news articles with sentence-level alignment (30,000 pairs of sentences) [14]

Machine Translation System

- *YakushiteNet* — Chinese-Japanese, English-Japanese and Japanese-English Machine Translation System [6]

Figure 1 shows the relation of the resources and the languages we used. Although the MT system is used only for unidirectional translation, the bilingual dictionaries and the parallel corpus are used for bidirectional translation.

There are two approaches for cross-lingual IR, the query translation approach and the document translation approach. The former translates queries to the language in which documents are written, and then monolingual IR is performed. Our system mainly uses the query translation approach.

Machine translation systems can be directly used for translating queries, but bilingual dictionaries and parallel corpora cannot. We use the cross-lingual PRF (CLPRF) method [2, 12] for query translation with bilingual dictionaries and parallel corpora. The CLPRF method is similar to the monolingual PRF method, but it uses parallel data. Given a query, the top M' documents are retrieved from the parallel data written in the source language, then N' terms are extracted from the corresponding parallel data written in the target language and used as a translated query. The method has two advantages. One is that the method can be applied to both bilingual dictionaries and parallel corpora. The other is that the method can handle both directions of translation regardless of the direction of the bilingual dictionary used in the translation; e.g., a Japanese-English dictionary can be used for both Japanese-to-English and English-to-Japanese translations. We also use monolingual PRF for queries before and after the query translation[1]. If there exist more than two resources for translation, we translate

[1]We used English documents of training (NTCIR-4) data for monolingual PRF of the E-C, E-J and E-K runs, because the English documents of the test data is not available at that time.

queries using each resource and finally combine the resulting translated queries.

Since we could not find any language resources for some language pairs, e.g. Japanese-Korean and Chinese-Korean, we use the pivot language method. For example, queries written in Japanese are translated to English queries first, and then the English queries are translated to Korean queries.

Table 1 shows the usage of the language resources in each run. Although the Chinese-English dictionary can be used for translating Chinese queries into English, we translate them by using Japanese as a pivot language because we have the Chinese-Japanese and Japanese-English MT system and it performed better than the bilingual dictionary. In the C-K run, we use two pivot languages; the Chinese queries are translated into Japanese first, then translated to English, and finally translated to Korean.

In the J-E run, we tried the document translation approach as well as the query translation approach. We translated all the English documents into Japanese using MT, and conducted J-J search against them.

2.5 Multilingual IR

In multilingual IR, we merge each result of SLIR and BLIR for the same query. Several merging methods for MLIR have been studied. The *round-robin* method interleaves the retrieved documents of different languages by assuming that the significance of the ranking in each language is equal. The *raw-score* method merges the retrieved documents using raw values of RSVs. The *normalized-score* method merges the retrieved documents using normalized values of RSVs. The normalized RSV of the ith ranked document, RSV_i^{norm}, is calculated from the original value RSV_i as follows:

$$RSV_i^{norm} = \frac{RSV_i - \min_j\{RSV_j\}}{\max_j\{RSV_j\} - \min_j\{RSV_j\}}. \quad (3)$$

The *Z-score* method [13] merges the retrieved documents using the Z-scores of RSVs calculated as follows:

$$RSV_i^Z = \frac{RSV_i - \mu}{\sigma}. \quad (4)$$

where μ is the average of RSVs and σ is the standard deviation.

In our system, RSVs are normalized by simply subtracting the average value:

$$RSV_i' = RSV_i - \mu. \quad (5)$$

We do not divide the values as the normalized-score and the Z-score methods because RSVs of our system are logarithms of (unnormalized) probabilities and this method performed well for the training data.

Run	MAP(% of SLIR)	Translation Direction (Method)
C-C	0.3330 (100%)	—
J-C	0.0779 (23%)	J-E(BD,PC,MT)+E-C(BD)
K-C	0.0377 (11%)	K-E(BD)+E-C(BD)
E-C	0.0853 (26%)	E-C(BD)
C-J	0.1932 (70%)	C-J(MT)
J-J	0.2763 (100%)	—
K-J	0.0583 (21%)	K-E(BD)+E-J(BD,PC)
E-J	0.1986 (72%)	E-J(BD,PC,MT)
C-K	0.1406 (32%)	C-J(MT)+J-E(BD,PC)+E-K(BD)
J-K	0.1612 (37%)	J-E(BD,PC,MT)+E-K(BD)
K-K	0.4334 (100%)	—
E-K	0.1171 (27%)	E-K(BD)
C-E	0.2356 (54%)	C-J(MT)+J-E(BD,PC)
J-E	0.3365 (77%)	J-E(BD,PC,MT)
K-E	0.1003 (23%)	K-E(BD)
E-E	0.4350 (100%)	—

Table 1. MAP of Formal Runs and Translation Methods (BD: Bilingual Dictionary, PC: Parallel Corpus, MT: Machine Translation; '+' means the use of pivot languages) (D-run, Rigid)

3 Evaluation Results and Discussion

In this section, we analyze our results of the NTCIR-5 formal runs.

3.1 Formal Run Results

The MAP values of formal runs are shown in Table 2 and Figure 2[2]. The performance of BLIR runs using Korean queries or documents is low, and the main reason seems to be the limited amount of the translation resources. The MAP values for cross-lingual IR are summarized in Table 1. Although the MAP values of BLIR in which MT is used are more than 70% of SLIR's performance, the values of BLIR in which only bilingual dictionary is used are less than 30% of that. The MAP values of PLIR varied from 11% to 54% of the SLIR's performance. Interestingly, the J-K run, which uses English as a pivot language and conducts Japanese-to-English translation then English-to-Korean translation, outperformed the E-K run which conduct only English-to-Korean translation, probably because the Japanese-to-English translation successfully expanded the queries.

3.1.1 Performance of J-J Run

Among the results of SLIR, only the J-J run's MAP value (Rigid) is lower than the average value of the participants. We conducted experiments with several settings in order to examine the reason, whether the

[2]Although we submitted two results for the J-E run; the result by the query translation method and the one by the document translation method, we refer to the one by the query translation if not mentioned otherwise.

	D-run		DN-run		T-run		TC-run		TDNC-run	
	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid	Relax	Rigid
C-C	0.4071	0.3330	0.4650	0.3994	0.4278	0.3584	0.4286	0.3422	0.4731	0.4031
J-C	0.0901	0.0779	0.1149	0.1092	0.1007	0.0912	0.1094	0.0907	0.1255	0.1135
K-C	0.0383	0.0377	0.0315	0.0268	0.0968	0.0929	0.0626	0.0477	0.0416	0.0324
E-C	0.0986	0.0853	0.1298	0.1106	0.1234	0.1113	0.1291	0.1159	0.1097	0.0981
C-J	0.2805	0.1932	0.3517	0.2665	0.2678	0.1757	0.3029	0.2053	0.3607	0.2641
J-J	0.3835	0.2763	0.5071	0.3847	0.4231	0.3083	0.4318	0.3090	0.5093	0.3865
K-J	0.0828	0.0583	0.0388	0.0368	0.1057	0.0813	0.0661	0.0415	0.0705	0.0530
E-J	0.2874	0.1986	0.4066	0.2976	0.2925	0.1970	0.3360	0.2393	0.3991	0.2970
C-K	0.1614	0.1406	0.1363	0.1291	0.1315	0.1168	0.1399	0.1262	0.1586	0.1380
J-K	0.1789	0.1612	0.1694	0.1544	0.1468	0.1309	0.1979	0.1730	0.1852	0.1681
K-K	0.4926	0.4334	0.5362	0.4776	0.4554	0.4033	0.5071	0.4492	0.5501	0.4940
E-K	0.1341	0.1171	0.1376	0.1219	0.1798	0.1528	0.1340	0.1124	0.1621	0.1393
C-E	0.2770	0.2356	0.2700	0.2346	0.2421	0.2033	0.2913	0.2536	0.2599	0.2229
J-E	0.3898	0.3365	0.4318	0.3766	0.4218	0.3679	—	—	—	—
K-E	0.1126	0.1003	0.0529	0.0381	0.1377	0.1180	0.1246	0.1023	0.1408	0.1204
E-E	0.4867	0.4350	0.5338	0.4781	0.4796	0.4239	0.4826	0.4102	0.5371	0.4766
C-CJKE	0.2259	0.1856	0.2557	0.2052	0.2026	0.1717	0.2349	0.1845	0.2496	0.1935
J-CJKE	0.1995	0.1706	0.2380	0.1890	0.2006	0.1771	0.2233	0.1825	0.2392	0.1854
K-CJKE	0.1030	0.0872	0.1138	0.1085	0.0937	0.0822	0.1171	0.1051	0.1462	0.1347
E-CJKE	0.1963	0.1522	0.2468	0.2110	0.2028	0.1596	0.2175	0.1753	0.2426	0.2100

Table 2. MAP of Formal Runs

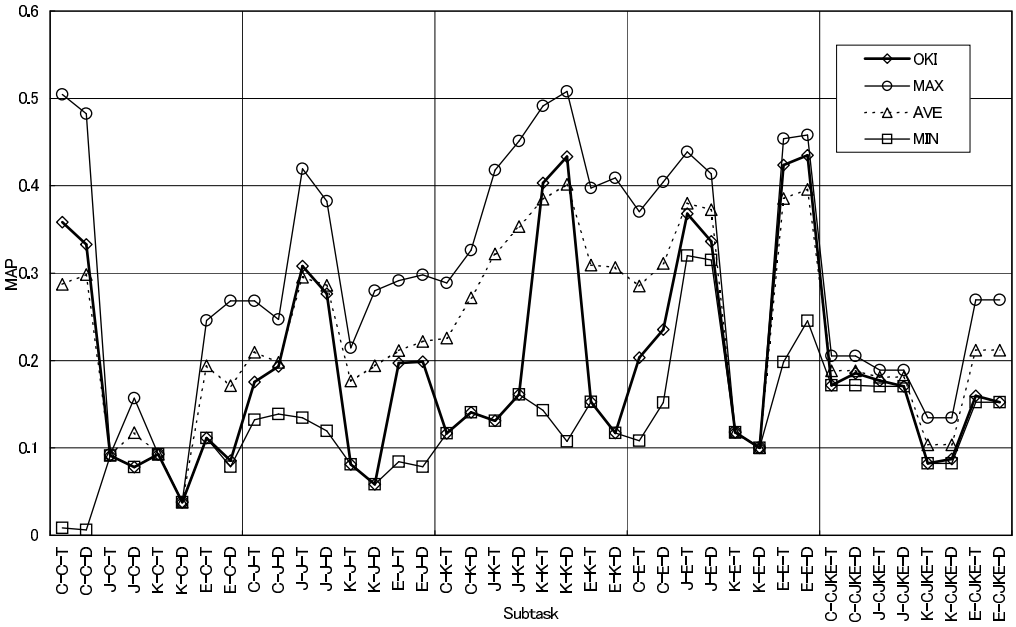

Figure 2. MAP of Formal Runs (T,D-run, Rigid)

Run	MAP (PRF Parameters)		
	Formal Run	Best(Training)	Best(Test)
C-C	0.3330 (M=10, N=10)	0.3493 (M=10, N=40)	0.3569 (M=20, N=40)
J-J	0.2763 (M=10, N=10)	0.3276 (M=30, N=30)	0.3528 (M=10, N=60)
K-K	0.4334 (M=10, N=10)	0.4390 (M=20, N=30)	0.4517 (M=30, N=30)
E-E	0.4350 (M=10, N=10)	0.4201 (M=30, N=10)	0.4380 (M=10, N=40)

Table 3. MAP for Different PRF Parameters (M: # of documents, N: # of terms) (D-run, Rigid)

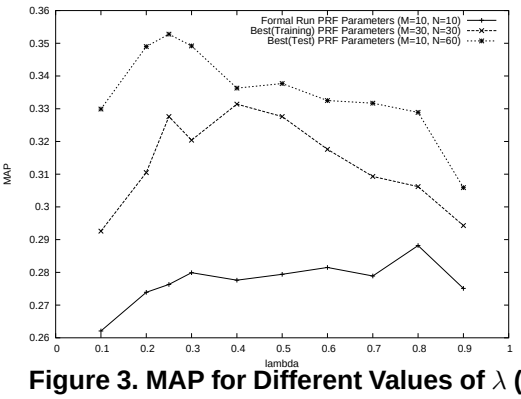

Figure 3. MAP for Different Values of λ (J-J-D-run, Rigid)

lower performance is essentially caused by our method or wrong parameter setting.

We varied parameters of PRF, the value M(the number of documents to be retrieved) and N(the number of terms to be extracted), to find the best parameters for the training (NTCIR-4) data and the test (formal run) data. Table 3 shows the MAP values on the test data with the parameters used in the formal run, the best parameters for the training data and the best parameters for the test data. The MAP value with the J-J formal run parameters is much smaller than that with the best parameters for the training data and the test data. We used the same PRF parameters ($M = 10, N = 10$) for all the runs and did not tune the parameters for each language in order to avoid overfitting to the training data. However, the inappropriate parameters decreased the performance of the J-J run. Figure 3 shows the MAP values for different values of λ(the smoothing parameter of the language models). Although we set the value to $\lambda = 0.25$ in our formal runs, the effect of the parameter was smaller than that of PRF parameters.

3.1.2 Methods and Resources for BLIR

In the J-E run, we conducted the query translation method and the document translation method. The document translation method uses the MT system, and the query translation method uses the bilingual dictionary, the parallel corpus and the MT system. To compare the performance with the different translation methods and resources, we conducted experiments with several settings.

Table 4 shows the results on the training data and the test data. In the experiments, the performance with query translation is better than that with document translation whether the monolingual PRF is used or not. From our investigation, the differences are mainly caused by the output of the MT systems (the query translation uses the Japanese-to-English MT and the document translation uses the English-to-Japanese MT). We used all the three language resources for query translation in the formal runs. Although the performance with the three resources was best on the training data, it was lower than that with only MT on the test data.

3.1.3 Merging Methods for MLIR

We conducted MLIR experiments to examine the performance for different merging strategies. Table 5 shows the results on the training data and the test data. Our system's method (Subtraction) had the highest MAP values for all the training data. However, on the test data, no single method achieved the highest MAP values for all the runs.

4 Conclusion

We developed the CLIR system which handles Chinese, Japanese, Korean and English, and participated in the SLIR, BLIR(PLIR) and MLIR subtasks. The system utilizes the bilingual dictionaries, the parallel corpus and the machine translation system for BLIR, and also uses the pivot language method for some language pairs. We submitted search results for all the 20 runs, however the performance of cross-lingual IR is not yet satisfactory. Adapting parameters of the language model-based IR and improving the performance of cross-lingual IR are left for future work.

Acknowledgements

We used CEDICT [3], ChaSen [7], EDICT [1], engdic [9], GETA [5] and Japanese-English News Article Alignment Data [14] in our research. We sincerely thank all the people who made these software and data.

This work was supported by a grant from the National Institute of Information and Communications Technology (NICT) of Japan.

References

[1] J. Breen. The EDICT Project, 2005. `http://www.csse.monash.edu.au/~jwb/edict.html`.

[2] J. G. Carbonell, Y. Yang, R. E. Frederking, R. D. Brown, Y. Geng, and D. Lee. Translingual Information Retrieval: A Comparative Evaluation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 708–715, 1997.

[3] P. Denisowski. CEDICT: Chinese-English Dictionary, 2005. `http://www.mandarintools.com/cedict.html`.

[4] D. Hiemstra. *Using Language Models for Information Retrieval*. Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[5] IPA. Generic Engine for Transposable Association (GETA), 2003. `http://geta.ex.nii.ac.jp/e/`.

[6] M. Kitamura and T. Murata. Practical Machine Translation System allowing Complex Patterns. In *Proceedings of the Ninth Machine Translation Summit*, pages 232–239, 2003. `http://www.yakushite.net/`.

[7] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. *Morphological Analysis System ChaSen version 2.2.8 Manual*. Nara Institute of Science and Technology, 2001.

[8] T. Nakagawa. Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 466–472, 2004.

[9] K. Paik and F. Bond. Enhancing an English/Korean Dictionary. In *Proceedings of Proceedings of Papillon 2003 Workshop*, 2003. (originally compiled by KwangSuk Lee).

Translation Method & Resources	MAP (J-E) (Training/Test)
Query Translation (Dictionary, Corpus, MT)	0.2841 / 0.3365
Document Translation (MT)	— / 0.3150
Query Translation (Corpus)	0.2189 / 0.2352
Query Translation (Dictionary)	0.1522 / 0.1845
Query Translation (MT)	0.2722 / 0.3593
Query Translation (MT, w/o monolingual PRF)	0.2102 / 0.2686
Document Translation (MT, w/o monolingual PRF)	— / 0.2571

Table 4. MAP of J-E run for Different Translation Methods and Resources (J-E-D-run, Rigid, Training / Test)

Merging Method	MAP (Training / Test)			
	C-CJKE	J-CJKE	K-CJKE	E-CJKE
Round-Robin	0.1001 / 0.1474	0.1229 / 0.1233	0.0818 / 0.0737	0.1183 / 0.1430
Raw-Score	0.0995 / 0.1716	0.1190 / 0.1574	0.0839 / 0.0585	0.1171 / 0.0880
Normalized-Score	0.1026 / 0.1600	0.1372 / 0.1349	0.0882 / 0.0758	0.1358 / 0.1568
Z-Score	0.1059 / 0.1655	0.1330 / 0.1408	0.0938 / 0.0934	0.1306 / 0.1592
Subtraction	0.1085 / 0.1856	0.1375 / 0.1706	0.0987 / 0.0872	0.1408 / 0.1522

Table 5. MAP of MLIR for Different Merging Method (D-run, Rigid, Training / Test)

[10] J. M. Ponte. *A Language Modeling Approach to Information Retrieval*. Ph.D. Thesis, Graduate School of the University of Massachusetts Amherst, 1998.

[11] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[12] M. Rogati and Y. Yang. Cross-Lingual Pseudo-Relevance Feedback Using a Comparable Corpus. In *CLEF 2001*, pages 151–157, 2001.

[13] J. Savoy. Report on CLIR Task for the NTCIR-4 Evaluation Campaign. In *Proceedings of the 4th NTCIR Workshop on Research in Informatin Access Technologies, Informatin Retrieval, Question Answering and Summarization*, 2004.

[14] M. Utiyama and H. Isahara. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 72–79, 2003. `http://www2.nict.go.jp/jt/a132/members/mutiyama/jea/`.

A Document Scoring using Language Models

RSVs by the $score_4$ model [4] is obtained as follows:

$$
\begin{aligned}
&RSV(d)\\
&= \log P(d|q),\\
&= \log P(d)P(q|d) - \log P(q),\\
&= \log P(d) + \log P(q|d) + c',\\
&\simeq \log P(d) + \log \prod_{t\in q} P(t|d)^{tf(t,q)} + c',\\
&\simeq \log P(d) + \log \prod_{t\in q} \{\lambda P'(t|d) + (1-\lambda)P'(t)\}^{tf(t,q)} + c',\\
&= \log P(d) + \log \prod_{t\in q\cap d} \{\lambda P'(t|d) + (1-\lambda)P'(t)\}^{tf(t,q)}\\
&\quad \prod_{t\in q-d} \{(1-\lambda)P'(t)\}^{tf(t,q)} + c',\\
&= \log P(d) + \log \prod_{t\in q\cap d} \left\{ \frac{\lambda P'(t|d)}{(1-\lambda)P'(t)} + 1 \right\}^{tf(t,q)}\\
&\quad \prod_{t\in q} \{(1-\lambda)P'(t)\}^{tf(t,q)} + c',\\
&= \log P(d) + \log \prod_{t\in q\cap d} \left\{ \frac{\lambda P'(t|d)}{(1-\lambda)P'(t)} + 1 \right\}^{tf(t,q)} + c'',\\
&= \log \frac{\sum_{t'} tf(t',d)}{\sum_{d'}\sum_{t'} tf(t',d')} +\\
&\quad \sum_{t\in q\cap d} \log \left\{ \frac{\lambda tf(t,d)\sum_{t'} df(t')}{(1-\lambda)df(t)\sum_{t'} tf(t',d)} + 1 \right\}^{tf(t,q)} + c'',\\
&= \log \sum_{t'} tf(t',d) +\\
&\quad \sum_{t\in q\cap d} tf(t,q) \log \left\{ \frac{\lambda tf(t,d)\sum_{t'} df(t')}{(1-\lambda)df(t)\sum_{t'} tf(t',d)} + 1 \right\} + c.
\end{aligned}
$$

where c', c'' and c are constants independent of d, and we assumed:

$$
P(d) = \frac{\sum_{t'} tf(t',d)}{\sum_{d'}\sum_{t'} tf(t',d')}.
$$

$$
P'(t|d) = \frac{tf(t,d)}{\sum_{t'} tf(t',d)}. \quad P'(t) = \frac{df(t)}{\sum_{t'} df(t')}.
$$