Balanced Query Methods for Improving OCR-Based Retrieval

Kareem Darwish
Electrical and Computer Engineering Dept.
University of Maryland, College Park
College Park, MD 20742
kareem@glue.umd.edu

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park
College Park, MD 20742
oard@glue.umd.edu

Abstract

Since many documents are available only print, improving OCR-based retrieval of scanned documents is an important problem. This paper presents a novel language independent technique for mapping queries from an error-free space to an OCR-degraded document space using a noisy channel model to produce possible degraded versions of query terms. The new technique yielded statistically significant improvements in retrieval effectiveness of as much as 39% over clean queries when tested on an Arabic document image collection.

1 Introduction

Information Retrieval (IR) is the process of satisfying a searcher's information need, expressed in the form of a query, by identifying documents that are likely to contain desired information. In many IR applications, queries and documents are represented in feature spaces that are related, but not identical. Well known examples include cross-language information retrieval (the process of finding documents in one language based on queries in a different language [4]), speech-based retrieval, and OCR-based retrieval. In such applications, a noisy channel model can be used to map between the query and document spaces, either mapping the queries into the document space or vice versa.

Much work has been done on correcting OCR errors in documents, which amounts to mapping the documents into the query space, in order to improve retrieval effectiveness (c.f., [7]). There has, however, been far less work on mapping queries into the document space based on the ways OCR would be likely to have misrecognized a term. This paper presents a novel technique for mapping queries from an error-free query space to a OCR-degraded document space using a noisy channel model for OCR degradation. The model is used to produce possible degraded versions of query terms to replace the original terms. The new technique is tested on an Arabic document image collection at three degradation levels.

2 Previous Work

All of the work that we are aware of on the problem of mapping query terms to OCR-degraded representations has adopted an approach based on term clustering. Harding et al. computed the q-gram distance for all the terms in an OCR-degraded collection to each of the query terms. Collection terms that we found to be "near" a query term were then treated as synonyms of that term when forming the query [2]. They used the InQuery synonym operator for this purpose, which separately computes the term frequency and the document frequency of a query term as if all of the synonyms were identical. With this approach, Harding et al. reported statistically significant improvement in mean average precision (a commonly reported retrieval effectiveness measure) of 12% over use of only the query term for English words with four relatively small test collections.

A similar approach was used by Hawking, again for English, in which each query term was mapped into all the terms in the degraded documents within some fixed edit distance [3]. The edit distance measure was constrained somewhat, with the set of possible substitutions constrained to a manually constructed set of the most likely character confusions. For example, the letter "o" could be substituted for "e," but not for "l." This approach resulted in a statistically significant improvement in retrieval effectiveness over the use of only the uncorrupted query terms on a relatively large collection (from TREC-4), again using English words.

3 Methodology

The experiments reported in this paper were conducted on a relatively small collection of Arabic text that we call "Zad," for which we have both scanned pages and electronic text as ground truth [1]. The collection is comprised of 2,730 documents extracted from *Zad Al-Me'ad*, a printed book for which an accurately character coded electronic version (the "clean text") is also available [5]. Three sets of OCR outputs for the same

documents were available: print resolution (300x300 dots per inch (dpi)) as originally scanned, and down-sampled versions that approximated fine fax resolution (200x200 dpi) and standard fax resolution (200x100 dpi). The test collection includes 25 written topic descriptions and associated relevance judgments. All diacritics (short vowels) were removed and character normalizations were performed as described by Darwish and Oard [1]. Three sets of experiments were run with different index terms: character 3-grams (3g), character 4-grams (4g), and lightly stemmed words (ls). Darwish and Oard found those index terms to be among the most effective for OCR-based retrieval of Arabic [1]. All the experiments were performed using PSE, a freely redistributable vector space system designed for experimental evaluation of information retrieval algorithms. PSE uses the well-known Okapi BM-25 formula to compute term weights [6].

OCR degradation was modeled with a position-sensitive unigram character distortion model trained on a sample of real OCR results and the corresponding clean text. Automatic alignment between the OCR-degraded text and the associated clean text was used to model a manual correction process, as might be used to learn character distortion models in real applications. Experiments were run with aligned sets of between 500 and 20,000 words in order to characterize the sensitivity of the techniques being evaluated to the amount of available training data. The appearance of Arabic characters varies with position, so distortion probabilities were separately modeled for beginning, middle, end, isolated characters.

Formally, given a clean word with characters $C_1..C_i..C_n$ and the resulting word after OCR degradation $D_1..D_j..D_m$, where D_j resulted from C_i, ε represents the null character, and L is the position of the letter in the word (beginning, middle, end, or isolated), the three edit operations for the models would be:

$$P_{\text{substitution}}(C_i \rightarrow D_j \mid L) = \frac{count(C_i \rightarrow D_j \mid L)}{count(C_i \mid L)}$$

$$P_{\text{deletion}}(C_i \rightarrow \varepsilon \mid L) = \frac{count(C_i \rightarrow \varepsilon \mid L)}{count(C_i \mid L)}$$

$$P_{\text{insertion}}(\varepsilon \rightarrow D_j \mid L) = \frac{count(\varepsilon \rightarrow D_j \mid L)}{count(C_i \mid L)}$$

If the count in the numerator was zero, the computation was repeated without conditioning on position.

A separate model was trained for each resolution (pring, fine fax, and standard fax). Two factors made automatic alignment of the OCR output to the clean text challenging. First, the printed and clean text versions in the Zad collection were obtained from different sources that exhibited minor differences (mostly substitution or deletion of particles such as *in*, *from*, *or*, and *then*). Second, some areas in the scanned images of the printed page exhibited image distortions that resulted in relatively long runs of OCR errors. The alignment was performed using SCLITE from the National Institute of Standards and Technology (NIST). SCLITE employs a dynamic programming string alignment algorithm that attempts to minimize the Levenshtein distance (edit distance) between two strings. Conceptually, the algorithm uses identical matches to anchor alignment, and then uses word position with respect to those anchors to estimate an optimal alignment on the remainder of the words.

SCLITE was originally developed for speech recognition applications, but in OCR applications additional character-level evidence is available. SCLITE alignments were therefore accepted only if the number of character edit operations was less than or equal to 50% of the length of the shorter of the two matched words. To align the words that were not aligned by SCLITE the following algorithm was used:

1. Using the existing alignments as anchors, given an unaligned word at position l from the preceding anchor in a clean document, sequentially compare it to the words, in the corresponding degraded document between the corresponding pair of anchors with position l' from the preceding anchor where $|l'-l| \leq 5$.
2. When comparing two words, if the difference between their respective word lengths was less than or equal to 2 characters and the number of edit operations between the two words (using Lenvenshtien's edit distance) was less than a certain percentage q of the word length of the shorter one (the percentage q was the number of edit operation divided by the length of the shorter word), then the newly aligned words were used as anchors. Initially, q was set to 60%.
3. Steps 1 and 2 were iterated two more times using the new anchors with q equal to 40% and 20% to attempt to find more alignments.

This alignment technique works well for the print and fine fax resolutions, but it is a significant source of errors for highly degraded cases (e.g., standard fax resolution).

Given a pair of aligned words, they were aligned at the character level by finding the edit distance between them using the Levenshtein edit distance algorithm and then back-tracing the algorithm to identify insertions, deletions, and substitutions.

Based on the training data, a garbler was built to read in a clean word $C_1..C_i..C_n$ and synthesize OCR degradation to produce n degraded versions of each query term $D'_1..D'_j..D'_m$, where n was varied between 1 and 50. The garbler was designed to produce "balanced queries" in which is more probable degraded versions would appear more often, and would thus have a greater

effect retrieval. For each garbled word, given a character C_i, the garbler chooses a random edit operation to perform (using a randomly seeded random number generator) based on the probability distribution for the possible edit operations for one of the models. If the chosen edit operation is insertion, the garbler picks a character to be inserted depending on the distribution of possible insertions. If the edit operation is a substitution, the model substitutes the character for another one based on the probability distributions of the possible substitutions.

The experiments were designed to answer the following questions:

1. Does mapping queries into document space improve retrieval effectiveness?
2. If so, how many training words are necessary to train the OCR-degradation model?
3. What is the optimal value of *n* (number of degraded versions for a query term)?

4 Results and Discussion

All the results were compared to a baseline generated using the query terms without any garbling. Baseline mean average precision results for the Zad collection are presented in Table 1. Tables 2, 3, and 4 at the end of the paper show the results for each contrastive condition. As can be seen, the method produced statistically significantly better results than the baseline for all index terms (based on a paired two-tailed *t*-test with $p<0.05$). Using more than 5,000 training words was found to offer little benefit for retrieval effectiveness, but smaller training sets yielded significantly worse results. This indicates that an adequate amount of training data could be hand corrected in just a few hours. Larger numbers of garbled terms (35 or 50) yielded the best results, indicating that the improved fidelity resulting from repeated sampling was helpful.

Table 1: Baseline Results for the Zad Collection.

	Mean Avg. Precision			
Index Term	Clean	Print	Fine Fax	Std Fax
3g	0.50	0.44	0.33	0.23
4g	0.53	0.46	0.32	0.22
ls	0.52	0.43	0.24	0.19

The maximum observed relative improvements over the clean-query baseline for the same combination of indexing term and resolution were as follows:

Resolution	Index Term	Maximum Relative Improvement
Print	3g	8%
	4g	7%
	ls	4%
Fine Fax	3g	13%
	4g	29%
	ls	36%
Standard Fax	3g	14%
	4g	27%
	ls	22%

Smoothing occurrence frequencies is often helpful when estimating distortion probabilities, so a second set of experiments was also tried with a simple variant of regression towards the most likely condition – in this case, the unchanged query. Document scores from the baseline runs for 3-grams, 4-grams, and lightly stemmed words were therefore combined with the document scores from the random garbling runs. When combining the scores, new document scores were:

$$\text{newScore} = \alpha * \text{baselineScore} + (1-\alpha) * \text{garbledScore}$$

Note that the smoothing here is applied to the effect (the document score) rather than directly to the probability distribution; this is a common approach to combining evidence in information retrieval applications. To find optimal values for α, baseline runs were combined with all previous garbled runs at values of α that varied between 0.1 and 0.9 with increments of 0.1. For each condition, the value of α that produced the highest mean average precision was noted. The average of the optimal value of α was computed separately for 10, 20, 35, and 50 garbled versions of each query term, and the resulting values were then averaged to provide a single value for use with each combination of indexing term and resolution. The resulting averages values of α were as follows:

Index Term	Print	Fine Fax	Standard Fax
3g	0.4	0.3	0.4
4g	0.2	0.2	0.5
Ls	0.5	0.4	0.5

The results for these combinations are summarized in Tables 5, 6, and 7. The maximum observed percent improvements over the clean-query baseline were for the same combination of indexing term and resolution were as follows:

Resolution	Index Term	Relative Improvement
Print	3g	7%
	4g	7%
	ls	4%
Fine Fax	3g	16%

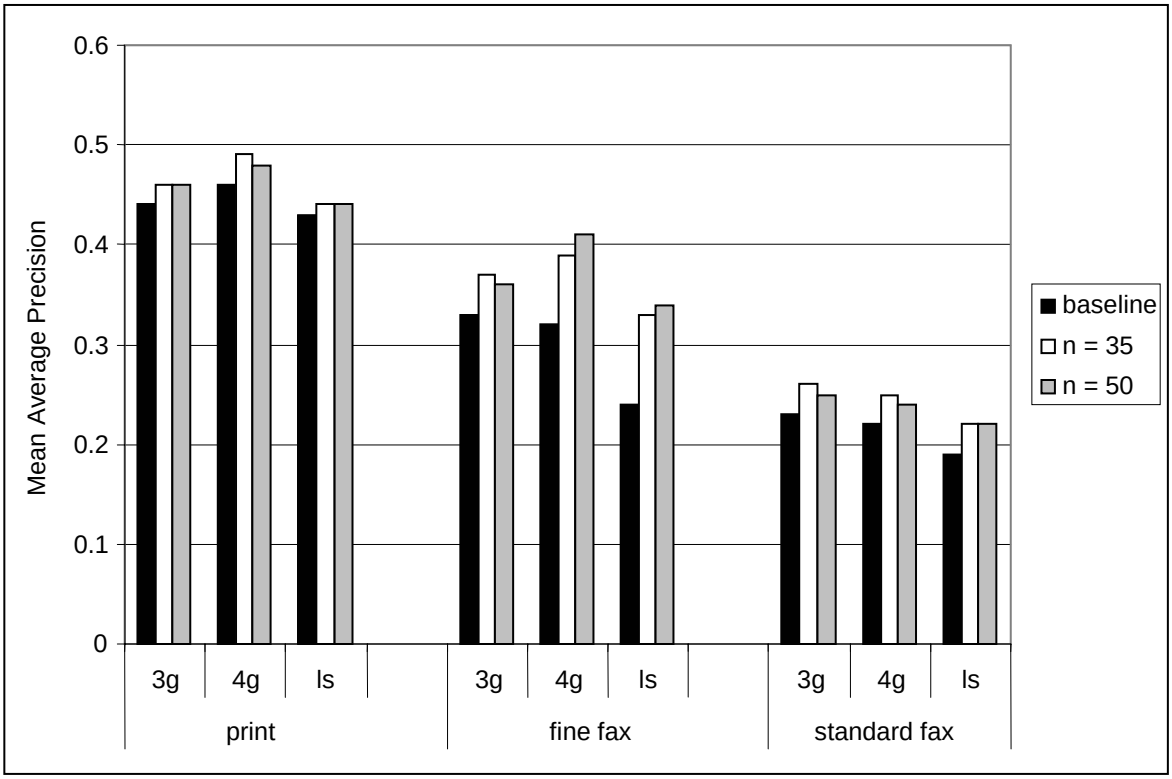

Figure 1: Combination of evidence results for different index terms using 5,000 training words

	4g	28%
	ls	39%
Standard Fax	3g	16%
	4g	16%
	ls	25%

Because the parameters were trained on the test collection, these should be thought of as upper bounds rather than as estimates of what can be accomplished in a real application. Little effect on the maximum relative improvement was observed, but a comparison of Tables 2-7 reveals that statistical significance was more often observed in the observed differences. Again, a greater number of garbled terms (35 or 50) yielded the best results, and providing more than 5,000 words of training data was not generally helpful.

Figure 1 shows combination of evidence results for different index terms and resolution with 5,000 training words. For print resolution, the maximum gains are modest, in part because the baseline results were excellent. Indeed, the mean average precision for the OCR-degraded collection using the mapping technique was statistically indistinguishable from retrieval of uncorrupted documents for 3-grams and 4-grams.

Balanced queries based on garbling showed the greatest relative improvements for fine fax resolution. Longer index terms (4-grams and lightly stemmed words) seemed to benefit the most, with a 39% improvement in retrieval effectiveness observed for lightly stemmed words.

For the standard fax resolution, the balanced query garbling technique often statistically outperformed the baseline. The observed relative improvements in retrieval effectiveness were smaller than those observed for fine fax resolution, perhaps because of some incorrect alignments in the training data (a factor that would not be present with hand-corrected training data).

4 Conclusions and Future Work

This paper presented a technique for mapping query terms into an OCR-degraded document space using a noisy channel model. The technique proved to be effective across a range of degradation levels, with relative improvements as high as 39% over a baseline in which no changes were made to the clean query. The technique has been demonstrated on a relatively small Arabic test collection, but the key ideas behind the technique are language independent. At present, the only large test collections for evaluation of retrieval effectiveness in OCR-degraded collections are based on synthetic (modeled) distortion of clean test collections, so a fair evaluation on a large collection will need to await the creation of a large collections based on actual OCR. The size and statistical significance of the observed improvements on the Zad collection do, however, suggest that benefits from this technique are likely to be seen on collections of any size.

There are two obvious ways in which the work reported in this paper could be extended. First, it would be worthwhile to explore the use of InQuery's weighted sum and synonym operators. Query term weighting can achieve a finer degree of granularity than the query term replication used in the experiments reported above, and the use of the synonym operator to implement "structured queries" is known to be helpful in similar applications such as cross-language retrieval. The experiments with model-based generation of plausible alternative expressions of query terms in an OCR-degraded collection reported in this paper suggest that those additional research directions would be well worth exploring.

References

[1] Darwish, K. and D. Oard. *Term Selection for Searching Printed Arabic*. SIGIR 2002, 261-268, 2002.

[2] Harding, S., W. Croft, and C. Weir. *Probabilistic Retrieval of OCR-degraded Text Using N-Grams*. European Conference on Digital Libraries, 1997.

[3] Hawking, D. *Document Retrieval in OCR-Scanned Text*. Sixth Parallel Computing Workshop, Kawasaki, Japan, November 1996.

[4] Oard, D. and B. Dorr, *A Survey of Multilingual Text Retrieval*. UMIACS, University of Maryland, College Park, 1996.

[5] Al-Areeb Electronic Publishers, LLC. 16013 Malcolm Dr., Laurel, MD 20707, USA.

[6] Robertson, S. and K. S. Jones. *Simple proven approaches to text retrieval*. Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997.

[7] Taghva, K., J. Borsack, and A. Condit. *An Expert System for Automatically Correcting OCR Output*. Proceedings of the SPIE - Document Recognition, pages 270--278, 1994.

Table 2: Print Resolution – All results are of Mean Average Precision. Grey cells indicate statistically worse results than the baseline and black cells indicate statistically better results than the baseline

Index Term	No. of Training words	baseline	1	5	10	20	35	50
3g	500	0.44	0.31	0.41	*0.45*	*0.44*	*0.42*	0.42
	1,000	0.44	0.40	0.44	0.45	*0.46*	0.44	0.45
	2,000	0.44	0.38	*0.44*	*0.46*	*0.45*	*0.45*	0.44
	5,000	0.44	0.32	*0.47*	*0.47*	*0.46*	*0.47*	0.44
	10,000	0.44	0.34	*0.46*	*0.45*	*0.47*	*0.45*	0.45
	20,000	0.44	0.39	*0.46*	***0.48***	*0.46*	*0.46*	*0.45*
4g	500	0.46	0.29	0.42	0.46	0.47	*0.47*	0.46
	1,000	0.46	0.40	0.45	0.47	*0.47*	0.48	*0.48*
	2,000	0.46	0.37	0.47	***0.49***	*0.48*	*0.48*	0.47
	5,000	0.46	0.32	*0.48*	*0.48*	***0.49***	***0.48***	*0.48*
	10,000	0.46	0.32	*0.48*	*0.48*	***0.49***	***0.49***	*0.48*
	20,000	0.46	0.39	*0.48*	***0.48***	*0.48*	***0.48***	***0.49***
ls	500	0.43	0.22	0.39	0.41	0.41	0.41	0.42
	1,000	0.43	0.28	0.40	0.42	0.43	0.43	0.43
	2,000	0.43	0.25	0.41	0.44	0.43	0.43	0.42
	5,000	0.43	0.21	0.41	0.43	0.43	0.43	0.44
	10,000	0.43	0.28	0.38	0.41	0.43	0.43	0.43
	20,000	0.43	0.26	0.44	0.43	0.45	0.44	0.43

Table 3: Fine Fax Resolution – All results are of Mean Average Precision. Grey cells indicate statistically worse results than the baseline and black cells indicate statistically better results than the baseline

Index Term	No. of Training words	baseline	1	5	10	20	35	50
3g	500	0.33	0.19	0.31	0.32	0.34	0.35	0.34
	1,000	0.33	0.20	0.31	0.34	0.35	0.37	0.36
	2,000	0.33	0.15	0.32	0.35	0.36	0.35	0.35
	5,000	0.33	0.14	0.29	0.32	**0.37**	0.36	0.35
	10,000	0.33	0.16	0.29	0.33	0.35	0.35	0.36
	20,000	0.33	0.21	0.30	0.34	0.35	0.36	0.36
4g	500	0.32	0.22	0.34	0.36	0.36	**0.38**	**0.39**
	1,000	0.32	0.20	0.33	0.35	**0.37**	**0.40**	**0.39**
	2,000	0.32	0.13	0.32	0.37	**0.37**	**0.40**	**0.39**
	5,000	0.32	0.13	0.29	0.36	**0.39**	**0.40**	***0.41***
	10,000	0.32	0.17	0.29	0.35	**0.39**	**0.40**	***0.41***
	20,000	0.32	0.23	0.31	0.35	0.37	***0.41***	***0.41***
ls	500	0.24	0.16	0.25	0.25	0.30	**0.30**	**0.32**
	1,000	0.24	0.17	0.23	0.26	0.29	**0.33**	**0.32**
	2,000	0.24	0.09	0.24	0.28	0.31	0.31	0.30
	5,000	0.24	0.07	0.23	0.26	0.31	0.30	0.32
	10,000	0.24	0.11	0.24	0.24	0.30	0.31	0.31
	20,000	0.24	0.17	0.23	0.24	0.28	0.30	**0.32**

Table 4: Standard Fax Resolution – All results are of Mean Average Precision. Grey cells indicate statistically worse results than the baseline and black cell indicate statistically better results than the baseline

Index Term	No. of Training words	baseline	1	5	10	20	35	50
3g	500	0.23	0.14	0.19	0.22	0.23	0.24	0.23
	1,000	0.23	0.09	0.22	0.20	0.21	0.24	0.24
	2,000	0.23	0.09	0.17	0.20	0.22	0.24	0.23
	5,000	0.23	0.12	0.18	0.21	0.24	0.26	0.22
	10,000	0.23	0.10	0.22	0.25	0.24	0.24	0.24
	20,000	0.23	0.10	0.18	0.19	0.22	0.24	0.24
4g	500	0.22	0.12	0.19	0.20	**0.25**	0.23	**0.25**
	1,000	0.22	0.10	0.21	0.22	0.23	**0.25**	**0.26**
	2,000	0.22	0.06	0.17	0.20	0.23	**0.25**	0.25
	5,000	0.22	0.13	0.21	0.23	0.24	**0.27**	0.25
	10,000	0.22	0.08	0.21	0.22	0.23	0.25	***0.27***
	20,000	0.22	0.10	0.19	0.19	0.23	0.25	*0.28*
ls	500	0.19	0.08	0.14	*0.17*	*0.18*	*0.20*	*0.22*
	1,000	0.19	0.07	*0.16*	*0.17*	*0.19*	*0.23*	***0.23***
	2,000	0.19	0.05	0.14	*0.15*	*0.18*	*0.21*	*0.21*
	5,000	0.19	0.10	*0.17*	*0.18*	*0.19*	*0.22*	*0.21*
	10,000	0.19	0.05	0.16	*0.17*	*0.19*	*0.20*	*0.22*
	20,000	0.19	0.07	0.15	*0.15*	*0.18*	*0.21*	*0.21*

Table 5: Print Resolution -- Combination of Evidence – All results are of Mean Average Precision. Black cell indicate statistically better results than the baseline

Index Term	No. of Training words	baseline	10	20	35	50
3g	500	0.44	0.45	*0.45*	0.44	0.44
	1,000	0.44	0.44	*0.46*	*0.45*	*0.46*
	2,000	0.44	*0.46*	*0.46*	*0.45*	*0.45*
	5,000	0.44	*0.46*	***0.46***	*0.46*	*0.46*
	10,000	0.44	*0.46*	***0.47***	*0.46*	*0.45*
	20,000	0.44	***0.47***	***0.47***	*0.47*	*0.47*
4g	500	0.46	0.46	0.47	*0.48*	0.47
	1,000	0.46	0.47	0.47	*0.48*	***0.49***
	2,000	0.46	***0.49***	*0.48*	*0.48*	*0.48*
	5,000	0.46	*0.48*	***0.49***	***0.49***	*0.48*
	10,000	0.46	***0.49***	***0.49***	***0.49***	***0.49***
	20,000	0.46	***0.49***	***0.49***	***0.49***	***0.49***
ls	500	0.43	0.43	0.43	0.43	0.43
	1,000	0.43	0.43	0.44	0.44	0.44
	2,000	0.43	0.44	0.44	0.44	0.44
	5,000	0.43	0.44	**0.44**	**0.44**	**0.44**
	10,000	0.43	0.44	0.44	0.44	0.44
	20,000	0.43	**0.44**	0.44	0.45	0.44

Table 6: Fine Fax Resolution -- Combination of Evidence – All results are of Mean Average Precision. Black cell indicate statistically better results than the baseline

Index Term	No. of Training words	baseline	10	20	35	50
3g	500	0.33	0.33	0.34	**0.35**	0.36
	1,000	0.33	0.34	0.36	**0.37**	**0.37**
	2,000	0.33	0.36	0.37	0.36	**0.38**
	5,000	0.33	0.34	**0.37**	0.37	0.36
	10,000	0.33	0.35	**0.37**	0.37	**0.39**
	20,000	0.33	0.35	0.36	**0.37**	**0.37**
4g	500	0.32	0.35	0.36	**0.36**	**0.39**
	1,000	0.32	0.34	0.36	**0.39**	**0.39**
	2,000	0.32	**0.37**	**0.38**	**0.40**	**0.39**
	5,000	0.32	0.36	**0.38**	**0.39**	***0.41***
	10,000	0.32	**0.38**	**0.38**	**0.40**	**0.40**
	20,000	0.32	0.37	**0.38**	**0.40**	***0.41***
ls	500	0.24	0.26	**0.31**	**0.29**	**0.33**
	1,000	0.24	0.29	0.30	**0.32**	**0.33**
	2,000	0.24	**0.30**	**0.32**	**0.32**	**0.32**
	5,000	0.24	**0.30**	**0.32**	**0.33**	**0.34**
	10,000	0.24	0.28	**0.31**	**0.32**	**0.33**
	20,000	0.24	0.28	0.30	**0.33**	**0.34**

Table 7: Standard Fax Resolution -- Combination of Evidence – All results are of Mean Average Precision. Black cell indicate statistically better results than the baseline

Index Term	No. of Training words	baseline	10	20	35	50
3g	500	0.23	0.24	0.24	0.25	0.26
	1,000	0.23	0.24	0.23	0.25	0.24
	2,000	0.23	0.23	0.24	0.25	0.25
	5,000	0.23	0.23	0.24	**0.26**	0.25
	10,000	0.23	0.26	0.25	0.25	0.26
	20,000	0.23	0.23	0.25	**0.26**	0.26
4g	500	0.22	0.22	**0.23**	**0.24**	**0.24**
	1,000	0.22	0.23	0.23	**0.24**	**0.25**
	2,000	0.22	0.22	0.23	**0.24**	**0.24**
	5,000	0.22	**0.23**	**0.24**	**0.25**	**0.24**
	10,000	0.22	**0.24**	0.23	**0.24**	**0.25**
	20,000	0.22	0.22	0.23	**0.25**	**0.25**
ls	500	0.19	*0.20*	*0.20*	*0.21*	***0.22***
	1,000	0.19	*0.21*	***0.21***	***0.23***	***0.23***
	2,000	0.19	*0.20*	***0.21***	***0.22***	***0.22***
	5,000	0.19	*0.21*	***0.21***	***0.22***	***0.22***
	10,000	0.19	*0.20*	*0.20*	*0.21*	***0.23***
	20,000	0.19	*0.19*	*0.20*	***0.23***	***0.22***