Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection

Gareth J. F. Jones, Declan Groves, Anna Khasin, Adenike Lam-Adesina, Bart Mellebeek and Andy Way

School of Computing, Dublin City University, Dublin 9, Ireland
email: {gjones,dgroves,akhasin,adenike,bart.mellebeek,away}@computing.dcu.ie

Abstract. For the CLEF 2004 ImageCLEF St Andrew's Collection task the Dublin City University group carried out three sets of experiments: standard cross-language information retrieval (CLIR) runs using topic translation via machine translation (MT), combination of this run with image matching results from the VIPER system, and a novel document rescoring approach based on automatic MT evaluation metrics. Our standard MT-based CLIR works well on this task. Encouragingly combination with image matching lists is also observed to produce small positive changes in the retrieval output. However, rescoring using the MT evaluation metrics in their current form significantly reduced retrieval effectiveness.

1 Introduction

Dublin City University's participation in the CLEF 2004 ImageCLEF St Andrew's collection task comprised three sets of experiments for Dutch, French, German, Italian and Spanish topic languages. First, we explored the application of our existing CLIR system used in previous CLEF workshops [1] with topic translation using three web-based translation resources. Second, the output from our standard CLIR system was combined with image matching results resulted provided by the track organisers, generated using the VIPER system. Finally, we explored a novel approach to rescoring the potentially relevant documents retrieved using our standard system based on automatic machine translation (MT) evaluation metrics.

This paper is organised as follows: Section 2 briefly outlines the details of our standard retrieval system, Section 3 gives results for our experiments using standard MT-based CLIR and combination with the provided VIPER system image retrieval output, Section 4 reports our results using MT evaluation metrics, and finally Section 5 concludes the paper.

2 CLIR Retrieval System

The basis of our experimental retrieval system is the City University research distribution version of the Okapi system, as used in our previous CLEF participation [1]. In this system documents and search topics are processed to remove stopwords from a list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming, and terms are further indexed using a small set of synonyms.

Terms are weighted using the standard BM25 weighting scheme and all runs use our summary-based pseudo relevance feedback (PRF) method [2]. The summary generation method combines Luhn's keyword cluster method, a title terms frequency method, a location/header method and a query-bias method from to form an overall significance score for each sentence. Sentences are ranked by significance score and the top ranked ones used to form a document summary. PRF expansion terms are selected from these summaries. Full details of this PRF method are given in [2].

3 Standard CLIR and Text-Image Combination Experimental Results

For all the experiments reported here the Okapi system parameters were selected using the training topics provided for the track. The parameter values were set as follows: $K1 = 1.0$ and $b = 0.5$ for baseline runs and $K1 = 1.5$ and $b = 0.6$ for PRF runs. The 20 top ranked PRF expansion terms from the summaries of the top 5 ranked documents were added to the baseline topic, with the top 20 ranked documents used to rank potential expansion terms for selection. The original topic terms were upweighted by a factor of 3.5 relative to terms introduced by PRF.

Topics were translated into English, the language of the documents, using the following web-based MT systems: Systran (`http://www.systransoft.com/`), SDL (`http://www.freetranslation.com/`) and InterTrans (`http://www.intertrans.com/`). Results are shown for precision at cutoffs of 5, 10, 15 and 20 documents, average precision and total number of relevant documents retrieved. The total number of relevant images available in the document collection for these topics is 829. Monolingual English results are shown for comparison for both baseline results without feedback and with the application of PRF. Baseline CLIR results are given for Systran topic translation, and for PRF and other conditions results are given the three separate topic translations and a merged union of the translated topics.

3.1 Baseline Runs

Table 1 shows baseline retrieval runs for monolingual English and Systran topic translation without application of PRF. The reduction in average precision for

Table 1. Baseline retrieval runs using Systran topic translation.

		English	Dutch	French	German	Italian	Spanish
Prec.	5 docs	0.664	0.464	0.488	0.536	0.408	0.384
	10 docs	0.624	0.424	0.488	0.536	0.396	0.416
	15 docs	0.587	0.405	0.451	0.501	0.384	0.381
	20 docs	0.552	0.384	0.418	0.440	0.378	0.362
Av Precision		0.545	0.384	0.427	0.464	0.402	0.383
% chg.		—	-29.5%	-21.7%	-14.9%	-26.2%	-29.7%
Rel. Ret.		774	698	631	695	606	654
chg. Rel. Ret.		—	-76	-143	-79	-168	-120

CLIR compared to monolingual IR varies between -15% and -30% with a variation in loss of total relevant documents retrieved of between -76 and -168. There is no clear correlation between the loss in average precision and relevant documents retrieved.

3.2 PRF Runs

The text annotations of the images are typically very short, comprising only a few sentences. In developing our PRF system for this retrieval task, we compared our summary-based approach, developed for use with newspaper document archives, with a standard PRF approach selecting terms from complete documents. For news documents collections our summary-based PRF method consistently outperforms a document-based approach [2]. Since the documents in the St Andrew's collection are so short, we felt it unlikely that use of document summaries would be useful for PRF. We were a little surprised to find that selecting terms from summaries of even these short documents when using the CLIR training topics worked better than the whole document approach. The summary-based approach is used for all PRF runs reported in this paper.

Table 2. Monolingual runs with application of PRF.

		Sentences		
		1S	2S	3S
Prec.	5 docs	0.608	0.600	0.608
	10 docs	0.640	0.620	0.644
	15 docs	0.608	0.592	0.619
	20 docs	0.554	0.550	0.560
Av Precision		0.524	0.546	0.545
Rel. Ret.		809	809	809

Table 2 shows monolingual feedback results for document summaries of the top ranked 1, 2 and 3 sentences. It can be seen that there is little change in

Table 3. Text retrieval runs with application of PRF.

			SDL	INT	ST	MG
Dutch (3S)	Prec.	5 docs	0.520	0.296	0.480	0.504
		10 docs	0.472	0.276	0.500	0.472
		15 docs	0.451	0.264	0.467	0.445
		20 docs	0.398	0.244	0.420	0.402
	Av Precision		0.398	0.273	0.432 (+12.5%)	0.421
	% chg. mono.		-27.0%	-49.9%	-20.7%	-22.8%
	Rel. Ret.		683	637	709 (+11)	791
	chg. Rel. Ret.		-126	-172	-100	-18
French (1S)	Prec.	5 docs	0.456	0.560	0.496	0.432
		10 docs	0.472	0.532	0.496	0.432
		15 docs	0.461	0.512	0.475	0.403
		20 docs	0.412	0.472	0.438	0.376
	Av Precision		0.409	0.466	0.431 (+0.9%)	0.399
	% chg. mono.		-21.9%	-11.1%	-17.7%	-23.9%
	Rel. Ret.		666	707	658 (+27)	695
	chg. Rel. Ret.		-143	-102	-151	-114
German (3S)	Prec.	5 docs	0.592	0.528	0.512	0.648
		10 docs	0.592	0.528	0.540	0.632
		15 docs	0.563	0.475	0.507	0.603
		20 docs	0.498	0.426	0.454	0.528
	Av Precision		0.501	0.468	0.474 (+2.6%)	0.531
	% chg. mono.		-8.1%	-14.1%	-13.0%	-2.6%
	Rel. Ret.		763	804	691 (-4)	804
	chg. Rel. Ret.		-46	-5	-118	-5
Italian (3S)	Prec.	5 docs	0.400	0.280	0.424	0.352
		10 docs	0.400	0.288	0.444	0.384
		15 docs	0.403	0.296	0.429	0.389
		20 docs	0.380	0.292	0.404	0.352
	Av Precision		0.366	0.288	0.438 (+9.0%)	0.351
	% chg. mono.		-30.2%	-47.2%	-19.6%	-35.6%
	Rel. Ret.		633	591	602 (-4)	639
	chg. Rel. Ret.		-176	-218	-207	-170
Spanish (2S)	Prec.	5 docs	0.472	0.312	0.440	0.432
		10 docs	0.484	0.316	0.460	0.448
		15 docs	0.475	0.312	0.445	0.432
		20 docs	0.430	0.290	0.400	0.388
	Av Precision		0.444	0.318	0.406 (+6.0%)	0.398
	% chg. mono.		-18.7%	-41.8%	-25.6%	-27.1%
	Rel. Ret.		767	666	649 (-5)	755
	chg. Rel. Ret.		-42	-143	-160	-54

Table 4. Text retrieval runs with application of PRF for revised Spanish topics.

			SDL	INT	ST	MG
Spanish (revised) (2S)	Prec.	5 docs	0.520	0.320	0.488	0.440
		10 docs	0.532	0.320	0.492	0.488
		15 docs	0.499	0.312	0.451	0.477
		20 docs	0.464	0.302	0.414	0.422
	Av Precision		0.472	0.312	0.410	0.446
	% chg. mono.		-13.6%	-42.9%	-24.9%	-18.2%
	Rel. Ret.		775	657	647	774
	chg. Rel. Ret.		-34	-152	-162	-35

average precision compared to the baseline result in Table 1. There is an improvement in total relevant documents retrieved, but a reduction in precision at cutoff 5. This results suggests that for monolingual retrieval with short documents and topics while recall can be improved and average precision maintained, retrieval accuracy problems may be introduced for documents retrieved at high rank in the baseline run.

Table 3 shows feedback results for each topic language with the three MT systems and the merged translated topics. Spanish results shown in Table 3 are for the original topic release. PRF results for the later released revised Spanish topics are shown in Table 4. Spanish results in later tables all relate to the original translated topics as used in Table 3. The number of sentences in the summary for each topic language is shown in the left column. This was selected for each language pairs using the training topics. The percentage differences shown here are relative to the monolingual results for PRF using the same number of summary sentences. For Systran, the difference relative to the baseline result shown in Table 1is shown adjacent to the average precision and total relevant retrieved results. Comparing the runs for Systran translated topics, we can see that PRF produces an improvement in average precision for each language pair. There is no clear trend for relevant document recall, there is small improvement for Dutch and French, and a small decrease for the others. Comparing retrieval effectiveness for the alternative topic translations, it can be seen that different systems produce the best average precision for different language pairs, although in general InterTrans is the least effective. Results for the merged topics are rather mixed. It was hoped that the increased term coverage would improve recall and aid precision; this does happen in some cases, but in others it reduces effectiveness. Further investigation is needed into specific success and failures to see if any general conclusions can be made.

The results in Table 3 show that the effect of merging the three separate topic translations is variable. An alternative method of combining multiple topic translations in a CLIR system is to combine the output of the three separate runs by merging the individual ranked lists in a process of data fusion. we have previously successfully used data fusion to combine the output of multiple topic translations in CLIR for news retrieval in CLEF 2001 [4]. Table 5 shows results of merging the output from our three separate translation runs with PRF using

Table 5. Data fusion retrieval runs with PRF.

		Dutch (3S)	French (1S)	German (3S)	Italian (3S)	Spanish (2S)
Prec.	5 docs	0.360	0.480	0.512	0.376	0.392
	10 docs	0.324	0.476	0.516	0.352	0.408
	15 docs	0.293	0.451	0.488	0.344	0.384
	20 docs	0.264	0.430	0.444	0.328	0.342
Av Precision		0.284	0.426	0.445	0.327	0.376
% chg.		-47.9%	-18.7%	-18.3%	-40.0%	-31.1%
Rel. Ret.		742	712	793	651	747
chg. Rel. Ret.		-67	-97	-16	-158	-62

Table 6. Retrieval results using image matching output from the VIPER system.

Prec.	5 docs	0.320
	10 docs	0.196
	15 docs	0.144
	20 docs	0.114
Av Precision		0.091
Rel. Ret.		142

a simple summation of the matching scores. It can be seen that these results are in all cases, except French, lower with respect to average precision than the merged translated topic results in Table 3. In most cases there is also a small reduction in the total number of relevant documents retrieved. The merged result for French in Table 3 is itself unusual, since the merged topic average precision is lower than that of any of the individual translations.

3.3 Text and Image Combination Runs

The St Andrew's documents are composed of images and text annotations as described in [3]. The search topics are similarly composed of a search image and text description. The experiments in the previous sections are based only on text retrieval. It is interesting to consider whether retrieval effectiveness might be improved by making use of the image data. The track organizers provided a set of image matching retrieval results generated using the VIPER image retrieval system. For each sample topic image, the top ranked 500 images from the test collection were provided together with their matching scores. Table 6 shows retrieval results using only these VIPER results. These results are not particularly good, only a small proportion of the relevant documents having been retrieved. However, the VIPER system does successfully retrieve a number of relevant images, and precision at high ranks is reasonable, indicating that where relevant images have been retrieved, this can be achieved with good precision. Probably the topic images match well with relevant images similar to themself, but fail to locate other relevant documents with rather different images. The

Table 7. Retrieval runs fusing PRF runs with standard VIPER image matching results.

			SDL	INT	ST	MG
Dutch	Prec.	5 docs	0.504	0.272	0.480	0.488
		10 docs	0.480	0.276	0.508	0.464
		15 docs	0.448	0.264	0.469	0.445
		20 docs	0.396	0.242	0.422	0.398
	Av Precision		0.394	0.273	0.433	0.419
	Rel. Ret.		638	637	709	791
French	Prec.	5 docs	0.464	0.552	0.488	0.416
		10 docs	0.472	0.520	0.496	0.428
		15 docs	0.456	0.512	0.472	0.405
		20 docs	0.414	0.470	0.436	0.380
	Av Precision		0.407	0.466	0.428	0.399
	Rel. Ret.		666	707	658	695
German	Prec.	5 docs	0.600	0.528	0.528	0.664
		10 docs	0.604	0.524	0.548	0.636
		15 docs	0.557	0.472	0.504	0.594
		20 docs	0.502	0.428	0.460	0.530
	Av Precision		0.501	0.467	0.474	0.532
	Rel. Ret.		763	804	691	804
Italian	Prec.	5 docs	0.400	0.280	0.424	0.344
		10 docs	0.400	0.288	0.440	0.392
		15 docs	0.400	0.304	0.427	0.387
		20 docs	0.380	0.290	0.402	0.352
	Av Precision		0.369	0.289	0.437	0.351
	Rel. Ret.		633	591	602	639
Spanish	Prec.	5 docs	0.464	0.296	0.432	0.424
		10 docs	0.472	0.324	0.452	0.444
		15 docs	0.461	0.299	0.429	0.432
		20 docs	0.428	0.288	0.392	0.384
	Av Precision		0.441	0.316	0.405	0.397
	Rel. Ret.		767	666	649	755

VIPER system was not tuned for this task, so these results form a lower bound on its potential effectiveness for this task.

We wanted to see if these image results could usefully be combined with out existing text results. Table 7 shows results for a simple sum data fusion combination of the matching score for the PRF runs shown in Table 3 and the VIPER runs provided by the track organizers. The merged score $merge\_score(j)$ of document j is formed as follows,

$$merge\_score(j) = 1.5 \times text\_score(j) + 1.3 \times VIPER\_score(j)$$

These combination constants were again selected using the training topics. The results in Table 7 are only slightly different from the PRF runs. However, some potentially important positives can be taken from this. First, in image retrieval it is often found that adding image matching information does not improve

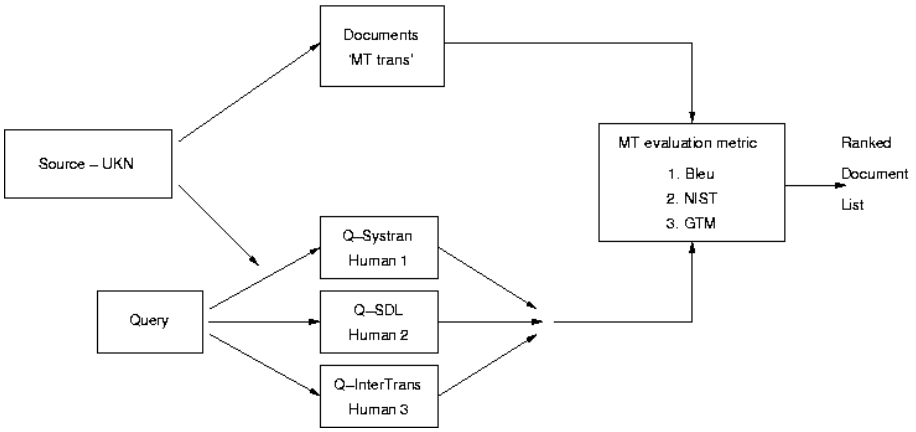

Fig. 1. Document scoring based on MT Evaluation metrics.

over text caption only retrieval. For our experiments in some cases the image matching score does help, albeit only marginally. Second, the VIPER system was not adjusted for the St Andrew's collection task, suggesting that a better image matching run should be possible with some task specific training of the image matching process. Given the relatively small number of relevant documents retrieved by the VIPER system, it is encouraging that it does not exert a significant negative impact on the text retrieval results.

4 Machine Translation Quality Metric Runs

For our final set of experiments we explored the use of a novel strategy for IR evaluation using automatic translation metrics. In recent years, several automatic MT evaluation methods have been proposed as a supplement to, or, in certain cases, a replacement for costly human MT evaluations [5][6][7][8]. These automatic evaluation methods rely on the idea that the quality of an MT output can be measured by its similarity to that of a professional human translation. With each of the currently available automatic evaluation methods, this similarity is measured using a word-error metric between the sentences in the MT-produced text output and the sentences in one or more human reference translations. The success of automatic MT evaluation depends largely on the amount of available comparable material and on the number of human reference translations, with more reference translations resulting in a more accurate measure of system performance.

In order to be able to use these metrics to calculate the similarity between a topic and a document in an IR system, we regard the original document and the MT-translated topic as translations of an unknown source text, as is shown in Figure 1.

In the first step, we extracted information from the headline and description sections of the original document text, as these sections contain the most relevant information pertaining to the documents. The same three sets of topic translations were used as in the previous experiments. The topic translations and documents were pre-processed to remove stopwords, capitalisation and punctuation. which allowed us to retain the most meaningful components of the text.

If we think of the topic translations as human reference translations, it is possible to measure the accuracy of the would-be 'machine translations' (the documents) using automatic MT evaluation metrics. The best 'machine translation' is the translation with the lowest word-error score with regard to the reference translations. The goal of our experiment was to find out to what extent the best 'machine translation' corresponded with a relevant document.

Experiments with the development topics showed that best results were obtained with a combination of 2 existing MT set evaluation methods (NIST and GTM) and an adaptation of the BLEU evaluation metric. BLEU ranks different MT output texts based a combination of an N-gram similarity score and a sentence brevity penalty with respect to a corpus of human reference translations. The BLEU evaluation script was adapted in two ways. First, we eliminated the sentence brevity penalty. The original BLUE metric penalizes short sentences to avoid the possibility that very short segments such as 'the' would receive a maximum score when compared to any sentence containing 'the'. This penalty is clearly not relevant for the retrieval task at hand. A second modification to the script consisted in allowing a non-zero BLEU score, regardless of the fact that, for one or more of the N-gram categories (unigram to 4-gram), no positive matches were found between the MT output and human reference translations.

The NIST metric differs from BLEU with respect to both the co-occurrence score and the sentence brevity penalty. The NIST metric alters the co-occurrence score in favour of lower order N-grams (i.e. low trigrams or quadrigram matches play less a role in the overall score) and more informative N-grams (i.e. N-grams that occur less frequently receive a higher weight). The sentence brevity penalty used by NIST is less severe than the one used by BLEU for sentences with small variations with respect to the reference translation.

GTM allows the calculation of standard precision and recall scores for automatically produced translations. It also calculates an f-measure score, which combines both the precision and recall scores for a given translation. It is this f-measure score, along with the NIST and adapted BLEU scores, that we used in our automatic ranking of the documents.

For our retrieval experiments the translated topics were ranked against the top 1000 documents retrieved for each topic using the text-only PRF approach described in the previous section. We used a summation of the NIST, f-measure and adapted BLEU scores to rescore the topic-document matches. We carried out two sets of experiments. In the first, we evaluated the retrieved document list against only one reference translation, as produced by each of the three online MT systems, giving us three resulting ranking lists of documents for each topic.

Table 8. Retrieval runs with pseudo relevance feedback.

			SDL	INT	ST	MG
Dutch	Prec.	5 docs	0.096	0.184	0.128	0.128
		10 docs	0.116	0.172	0.124	0.140
		15 docs	0.101	0.168	0.123	0.123
		20 docs	0.100	0.154	0.120	0.114
	Av Precision		0.105	0.127	0.141	0.121
	Rel. Ret.		638	637	709	791
French	Prec.	5 docs	0.104	0.128	0.096	0.088
		10 docs	0.128	0.120	0.128	0.112
		15 docs	0.133	0.115	0.125	0.101
		20 docs	0.130	0.106	0.122	0.100
	Av Precision		0.107	0.110	0.117	0.100
	Rel. Ret.		666	707	658	695
German	Prec.	5 docs	0.160	0.208	0.120	0.184
		10 docs	0.164	0.172	0.124	0.148
		15 docs	0.155	0.189	0.128	0.168
		20 docs	0.144	0.190	0.128	0.178
	Av Precision		0.146	0.169	0.132	0.148
	Rel. Ret.		763	804	691	804
Italian	Prec.	5 docs	0.168	0.128	0.160	0.136
		10 docs	0.132	0.132	0.140	0.112
		15 docs	0.133	0.139	0.128	0.109
		20 docs	0.126	0.128	0.120	0.104
	Av Precision		0.132	0.119	0.118	0.108
	Rel. Ret.		633	591	602	639
Spanish	Prec.	5 docs	0.128	0.120	0.128	0.128
		10 docs	0.140	0.140	0.140	0.132
		15 docs	0.157	0.128	0.149	0.133
		20 docs	0.160	0.128	0.156	0.131
	Av Precision		0.145	0.111	0.128	0.131
	Rel. Ret.		767	666	649	755

In the second set of experiments we merged the translated topics, using the three different translations of the topic as three different reference translations.

Table 8 shows results of document rescoring using MT evaluation metrics. Comparing these results to those using standard PRF methods in the earlier tables, it can be seen that the MT evaluation metrics are not effective for IR scoring in their present form. The main goal of our experiments was not to substantially improve the best available Image Retrieval methods, but to investigate a novel idea for IR of treating topic documents and translated user topics as comparable translations of an unknown source text. Clearly based on the results shown here we need to explore further whether this approach can be adapted successfully for IR applications.

5 Conclusions and Further Work

Our experiments for the CLEF 2004 ImageCLEF have demonstrated that our standard CLIR method works effectively for the short text documents in the St Andrew's collection, and further that there is potential for improvement in retrieval effectiveness from the use of image matching in cross-language image retrieval. Experiments using MT evaluation metrics for scoring CLIR have so far not been successful, but we intend to explore alternative means of applying this idea to see whether it can be used usefully in CLIR.

References

[1] A. M. Lam-Adesina and G. J. F. Jones. Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval. In *Proceedings of Fourth Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, C. Peters, J. Gonzalo, M. Braschler and M. Kluck (Eds.), Lecture Notes in Computer Science, Springer, Heidelberg, Germany, pages 271-285, 2004.

[2] A. M. Lam-Adesina and G. J. F. Jones. Applying Summarization Techniques for Term Selection in Relevance Feedback. In *Proceedings of the 24th Annual International ACM SIGIR*, pages 1-9, New Orleans, ACM, 2001.

[3] P. Clough, H. Müller and M. Sanderson. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In *Proceedings of the Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck and B. Magnini (Eds), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany (in print), 2005.

[4] G.J.F. Jones and A.M. Lam-Adesina. Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In *Proceedings of the Second Workshop of the Cross-Language Information Evaluation (CLEF 2001)*, C. Peters, M. Braschler, J. Gonzalo and M. Kluck (Eds.), Lecture Notes in Computer Science, Springer, Heidelberg, Germany, pages 59-77, Darmstadt, Germany, 2001.

[5] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318, Philadelphia, USA, 2002.

[6] G. Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Human Language Technology: Notebook Proceedings: 128-132. San Diego, 2002.

[7] General Text Matcher `http://nlp.cs.nyu.edu/GTM/`

[8] NIST's MT Evaluation Toolkit `http://www.nist.gov/speech/tests/mt/resources/scoring.htm`