Choosing the Best MT Programs for CLIR Purposes – Can MT Metrics Be Helpful?

Kimmo Kettunen

Department of Information Studies, University of Tampere, Finland
Kimmo.Kettunen@uta.fi

Abstract. This paper describes usage of MT metrics in choosing the best candidates for MT-based query translation resources. Our main metrics is METEOR, but we also use NIST and BLEU. Language pair of our evaluation is English → German, because MT metrics still do not offer very many language pairs for comparison. We evaluated translations of CLEF 2003 topics of four different MT programs with MT metrics and compare the metrics evaluation results to results of CLIR runs. Our results show, that for long topics the correlations between achieved MAPs and MT metrics is high (0.85-0.94), and for short topics lower but still clear (0.63-0.72). Overall it seems that MT metrics can easily distinguish the worst MT programs from the best ones, but smaller differences are not so clearly shown. Some of the intrinsic properties of MT metrics do not also suit for CLIR resource evaluation purposes, because some properties of translation metrics, especially evaluation of word order, are not significant in CLIR.

1 Introduction

Cross Language Information Retrieval (CLIR) has become one of the research areas in information retrieval during the last 10+ years [1]. The development of WWW has been one of the key factors that has increased interest in retrieval tasks where the language of the queries is other than that of the retrieved documents. One of the practices of CLIR has been translation of queries, or user's search requests. A popular approach for query translation has been usage of ready-made machine translation (MT) programs. As machine translation programs have been more readily available during the last years, and their quality has also become better, they are good candidates for query translation. Many of the programs are available as free web services with some restrictions on the number of words to be translated, and many standalone workstation programs can be obtained with evaluation licenses. CLIR can also be considered a good application area for "crummy MT", as Church and Hovy state it [2].

CLIR results for the languages give indirect evidence of the quality of machine translation programs used. It is evident that the better the query results are, the better the translation program, or translation resource in general, is. This was shown experimentally in McNamee and Mayfield [3, also 4] with purported degradation of translations on lexical level. Zhu and Wang [5] tested effects of rule and lexical degradation of a MT system separately and found that retrieval effectiveness correlated highly with

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 706–712, 2009.

the translation quality of the queries. Retrieval effectiveness was shown to be more sensitive to the size of the dictionary than the size of the rule base especially with title queries. Authors used NIST score as the evaluation measure for translation quality. Kishida [6] shows with a regressive model, that both ease of search of a given query and translation quality can explain about 60 % of the variation in CLIR performance.

In this paper we reverse the question: if we have several available MT programs, is it reasonable to test translation results of all of them in the actual query system or will MT metrics evaluation give enough basis for choosing the best candidates for further evaluation in the query system? This type of "prediction capability" may be useful, when there are lots of available MT systems for CLIR purposes for a language pair. It is not reasonable to test e.g. ten different query translations in the final CLIR environment, if the translation metrics will show the quality of the query translations with reasonable accuracy and thus predict also which MT systems will achieve best retrieval results.

2 Research Setting and Results

Kettunen [7] describes CLIR results of three languages, Finnish, German and Swedish with CLEF 2003 materials in Lemur query system. Four MT programs were used for query translation from English to German: Google Translate Beta, Babelfish, Promt Reverso and Translate It!

For better understanding of the translation quality of MT programs we did further evaluation of the German translation results of different MT systems with a machine translation evaluation metric METEOR 0.6 [8, 9, 10]. METEOR is based on a BLEU [11] like evaluation idea: output of the MT program is compared to a given reference translation, which is usually a human translation. METEOR's most significant difference to BLEU like systems is, that it emphasizes more recall than precision of translations [12]. The evaluation metric was run with exact match, where translations are compared to reference translation as such. Basically "METEOR evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a given reference translation". When "given a pair of strings to be compared, METEOR creates a word alignment between the two strings. An alignment is a mapping between words, such that every word in each string maps to most one word in the other string. This alignment is incrementally produced by a sequence of word-mapping modules. The 'exact' module maps two words if they are exactly the same." [13].

In our case the reference translation was the official CLEF 2003 translation of the English topics into German[1]. Four topics that do not have relevant documents in the collection were omitted from the test set, and the total number of topics was thus 56. Translations were evaluated in our tests topic by topic, i.e. each topic translation was a segment to be evaluated, and an overall figure for all the topic translations is given. Translations of title queries (T) were done separately from title and description queries (TD). Table 1 shows the results of METEOR's evaluations for all the English → German title and description MT outputs in their raw form. Table 2 shows results for title translation evaluations.

[1] If this methodology were to be used with e.g. web retrieval, where no known topic set and its translation is available, a test bed of "typical" queries and their ideal translations should be first established.

Table 1. Results of METEOR translation evaluation for long German topics

Metrics	Google Translate Beta	Babelfish	Promt Reverso	Translate It!
Overall system score	0.32	0.26	0.24	0.19
Matches	656	591	556	511
Chunks	329	326	305	314
HypLength	1101	1126	1083	1117
RefLength	1050	1050	1050	1050
Precision	0.60	0.52	0.51	0.46
Recall	0.62	0.56	0.53	0.49
1-Factor	0.61	0.54	0.52	0.47
Fmean	0.62	0.56	0.53	0.49
Penalty	0.49	0.54	0.54	0.60
Fragmentation	0.50	0.55	0.55	0.61
Number of segments scored	56	56	56	56

Table 2. Results of METEOR translation evaluation for short German topics

Metrics	Google Translate Beta	Babelfish	Promt Reverso	Translate It!
Overall system score	0.29	0.33	0.20	0.22
Matches	160	161	144	138
Chunks	91	82	96	86
HypLength	260	254	253	266
RefLength	244	244	244	244
Precision	0.62	0.63	0.57	0.52
Recall	0.66	0.66	0.59	0.57
1-Factor	0.63	0.65	0.58	0.54
Fmean	0.65	0.66	0.59	0.56
Penalty	0.56	0.50	0.65	0.61
Fragmentation	0.57	0.51	0.67	0.62
Number of segments scored	56	56	56	56

The meanings of the most important metrics in Tables 1 and 2 (bolded in tables) are as follows:

- *Overall system score* gives a combined figure for the result. It is computed as follows [9]: Score = Fmean * (1- Penalty).
- (Unigram) *Precision* = unigram precision is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the system translation.
- (Unigram) *Recall* = unigram recall is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the reference translation.
- *Fmean:* precision and recall are combined via harmonic mean that places most of the weight on recall. The present formulation of Fmean is stated in Lavie and Agarwal [12] as follows:
  Fmean = P * R/ α * P + (1- α) * R.
- *Penalty.* This figure takes into account the extent to which the matched unigrams in the two strings are in the same word order.

If we now compare the retrieval results of plain translated title and description queries reproduced from Table 4 in Kettunen [7] as Table 3, we notice that MAPs of the long query runs are in the order Google > Babelfish > Promt > Translate It! just as shown by the METEOR results in Table 1 by all the most important scores. For comparison, we also give MT metric scores from NIST [14] and BLEU metrics, given by *mteval-v.11b* [15].

Table 3. Mean average precisions of translated plain German TD queries and MT metrics scores. Metrics: M = METEOR, N = NIST, B = BLEU.

	MAP of TD queries	**Translation's quality score**		
		M	**N**	**B**
Google Translate Beta	39.9	0.32	4.8	0.26
Babelfish MT progam	30.3	0.26	4.2	0.19
Promt Reverso MT program	27.5	0.24	4.4	0.22
Translate It! MT program	26.1	0.19	3.8	0.17

Google's translation for whole topics is evaluated by far the best by METEOR and its MAP is also 9.6 % better than that of the next one. Babelfish and Promt are given more equal scores by METEOR, and their MAPs have only about a 3 % difference. Translation of Translate It! is evaluated clearly the worst by METEOR, and it gets the worst MAPs, although not very much inferior to Promt. Thus the overall quality of translation of whole topics seems to correlate with MAP of the retrieval. Correlation for MAPs of TD queries and METEOR scores is high: **0.94**. NIST metric's correlation to MAPs of TD queries is **0.85** and BLEU's correlation **0.85**.

Table 4 gives results of T queries from Kettunen [7] and relates MAPs of different MT systems to MT metrics.

Table 4. Mean average precisions of translated plain German T queries and MT metrics scores. Metrics: M = **METEOR**, N = NIST, B = BLEU.

	MAP of T queries	**Translation's quality score**		
		M	**N**	**B**
Google Translate Beta	30.1	0.29	3.1	0.28
Babelfish MT progam	24.2	0.33	3.2	0.31
Promt Reverso MT program	21.4	0.20	2.8	0.20
Translate It! MT program	20.5	0.22	2.5	0.18

METEOR's evaluation results of short queries differ from the MAP order. METEOR gives the order Babelfish > Google > Translate it! > Promt, while order by MAPs is Google > Babelfish > Promt > Translate It! MAPs of Promt and Translate It! do not differ much, and neither do their overall METEOR scores. But Google's MAP is much better than Babelfish's, so the METEOR result for title translations is confusing. A closer examination of the figures in Table 2 reveals that Google's penalty score with T queries is much higher than Babelfish's. Penalty scores word order of translations giving a lower score when the translation's word order is closer to the

reference's word order. It is apparent that the difference in the overall system score is due to the differences in penalty score, as other scores of Google are quite close to Babelfish's. Word order of translations is relevant from a translation point of view but it does not affect IR results, so this should be taken into account when using the METEOR metric. Effect of *Penalty* should either be discarded wholly or minimized somehow. If this is taken into account, METEOR was also able to clearly indicate the two best title translations and two worst title translations, although the order of evaluation results differed from the retrieval result order due to metric's inner logic. Correlation for T query MAPs and METEOR scores is lower than for TD queries**, 0.63**. Correlation for T query MAPs of NIST metric is **0.72** and BLEU's **0.71.**

3 Discussion

Our purpose in this research was to show the impact of the quality of MT to CLIR performance and thus make it possible to use MT metrics results as a prediction of translated queries' performance. It is self-evident that the quality of the translation affects results of retrieval, but the most important factor in query translation is the choice of vocabulary, not any other aspect of translation quality, e.g. word order of translations does not affect IR results [4]. We evaluated English → German translations mainly with one automatic MT evaluation program, METEOR 0.6, and got results that were mostly in accordance with the retrieval results: the MT program that got clearly the best evaluation scores from METEOR with whole topics was also clearly the best performer in CLIR evaluation. Other programs were also evaluated in the same order by METEOR as they performed in retrieval runs, but the differences of MT evaluation scores were perhaps not that clear as the CLIR performance differences. With titles of the topics the results of translation evaluation were more problematic: the best IR performer, Google's Translate, was evaluated the second best translation by METEOR, but this was due to the inner logic of the metrics, that also evaluates word order of translations. Overall it seems that evaluation scores of a MT metric give a fair indication of retrieval results, but the use of MT metrics would need more evaluation in this use. MAPs of retrieval and scores given by metrics correlate clearly, but different metrics also give slightly different quality scores for translations of different systems. In clearest cases (best vs. worst) the scores given by metrics indicate clearly also MAP results, but when differences in scores are small, evaluation is not that indicative. We suggest that use of a MT metric in CLIR translation resource evaluation can be beneficial in following aspects: it is easier to evaluate capabilities of several possible MT systems first with MT metrics to screen out the worst candidates and proceed after that to normal query result evaluation with fewer systems to pick the best one for the specific query translation task at hand. This was shown also in Kettunen [16], where 12 different En → De MT programs were evaluated with METEOR. Worst and best performers (by MAP) were clearly shown already by METEOR scores and correlations between MAPs and MT quality scores were **0.86** with TD queries and **0.61** with T queries. It would also be beneficial, if MT metrics could be fine-tuned for CLIR resource evaluation use by omitting weighting of word order of translations, which is not relevant in this use. Perhaps also some other fine-tuning would be needed for MT metrics in this specific use. Also the

impact of varied translations should be further studied with possibly more reference translations as is done in MT metrics evaluation connected to machine translation system evaluation The effect of different language pairs could also be further studied, although Clough and Sanderson [17] find a clear correlation of MAPs and MT quality scores for translations from six source languages to English.

Acknowledgements

This work was supported by the Academy of Finland grant number 1124131.

References

[1] Kishida, K.: Technical Issues of Cross-Language Information Retrieval: A Review. Information Processing & Management 41, 433–455 (2005)
[2] Church, K.W., Hovy, E.H.: Good applications for crummy machine translation. Machine Translation 8, 239–258 (1993)
[3] McNamee, P., Mayfield, J.: Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In: Proceedings of Sigir 2002, Tampere, Finland, pp. 159–166 (2002)
[4] Kraaij, W.: TNO at CLEF-2001: Comparing Translation Resources. In: Working Notes for the CLEF 2001 Workshop (2001), http://www.ercim.org/publication/ws-proceedings/CLEF2/kraaij.pdf
[5] Zhu, J., Wang, H.: The Effect of Translation Quality in MT-Based Cross-Language Information Retrieval. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th annual Meeting of the ACL, pp. 593–600 (2006)
[6] Kishida, K.: Prediction of performance of cross-language information retrieval system using automatic evaluation of translation. Library & Information Science Research 30, 138–144 (2008)
[7] Kettunen, K.: MT-based query translation CLIR meets Frequent Case Generation (submitted)
[8] Lavie, A., Agarwal, A.: The METEOR Automatic Machine Translation Evaluation System, http://www.cs.cmu.edu/~alavie/METEOR/
[9] Banerjee, S., Lavie, A.: METEOR: Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, pp. 65–72 (2005)
[10] Banerjee, S., Lavie, A.: METEOR: Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, pp. 65–72 (2005)
[11] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311–318 (2002)
[12] Lavie, A., Sagae, K., Jayarman, S.: The Significance of Recall in Automatic Metrics for MT Evaluation. In: Frederking, R.E., Taylor, K.B. (eds.) AMTA 2004. LNCS, vol. 3265, pp. 134–143. Springer, Heidelberg (2004)

[13] Lavie, A., Agarwal, A.: METEOR: An automatic Metric for MT Evaluation with High Levels of Correlation with Human judgements. In: Proceedings of the Second Workshop on Statistical Machine Translation, Prague, June 2007, pp. 228–231 (2007)
[14] Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145 (2002)
[15] Mteval-v.11b, http://www.nist.gov/speech/tools/
[16] Kettunen, K.: Facing the machine translation Babel in CLIR – can MT metrics help in choosing CLIR resources? (2008) (manuscript)
[17] Clough, P., Sanderson, M.: Assessing Translation Quality for Cross Language Image Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 594–610. Springer, Heidelberg (2004)