Recent Advances in Intelligent Information Systems
ISBN 978-83-60434-59-8, pages 103–116

Facing The Machine Translation Babel in CLIR – Can MT Metrics Help in Choosing CLIR Resources?

Kimmo Kettunen

Kymenlaakso University of Applied Sciences, Finland
kimmo.kettunen@kyamk.fi

Abstract

This paper describes usage of MT metrics in choosing the best candidates for MT-based query translation resources in Cross-Language Information Retrieval. Our metrics is METEOR. Language pair of our evaluation is English → German, because METEOR metrics does not offer very many language pairs for comparison. English → German has also available many MT programs that can be used in evaluation. We evaluated translations of CLEF 2003 topics of twelve different MT programs with MT metrics and compare the metrics evaluation results to mean average precision results of CLIR runs. Our results show, that for long topics the correlations between achieved MAPs and MT metrics are high (0.88), and for short topics lower but still clear (0.59). Overall it seems that METEOR can easily distinguish the worst MT programs from the best ones, but smaller differences are not so clearly seen. Some of the intrinsic properties of METEOR metrics do not also suit for CLIR resource evaluation purposes, because some properties of the translation metrics, especially evaluation of word order, are not significant for CLIR resource evaluation.

1 Introduction

Cross Language Information Retrieval (CLIR) has become one of the research areas in information retrieval during the last 10+ years (Kishida, 2005). 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 for most common (European) languages, 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 (1993) state it.

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 (2002; cf. also Kraaij, 2001) with purported degradation of translations on lexical level. Zhu and Wang (2006) tested effects of rule and lexical degradation of a MT system separately and found that retrieval effectiveness correlated highly with 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 (2008) 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 partly reverse the question: if we have several available MT programs for a language pair, is it reasonable to test translation results of all of them in the actual query system or will MT metrics evaluation results give enough basis for choosing the best candidates for further evaluation in the query system? This kind 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 sets of 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

We evaluated En → De translations of CLEF (Cross-Language Evaluation Forum) 2003 topics with twelve MT programs in Lemur query system (http://www.lemurproject.org/). The used MT programs were Google Translate Beta, Babelfish, Promt Reverso, Systran, IBM WebSphere, LEC Translate2Go, SDL Enterprise Translation Server, Translate It!, InterTran, Translated, Hypertrans and MZ-Win Translator. Most of the translation programs were available either as free trial versions or as web services. If the web service had limitations in the number of words to be translated, the topic set was split to smaller chunks. We translated separately title and title and description parts of the topics. Topic numbers and XML tags were omitted from the topics before translation. All T (title) and TD (title and description) topics ended in a full stop when given as input for the MT systems, and TD topics had more than one sentence or sentence like construction. The mean length of the original English TD parts of topics is 18.8 words and the mean length of T parts is 3.7 words. Mean length of the reference translations of TDs is 17.25 words, and for Ts 3.15 words. Table 1 lists the MT programs and their sources.

After translation we ran all the query translations in the Lemur query system with German CLEF 2003 collection and got CLIR query evaluation results from *trec.eval* as mean average precision (MAP) figures (per cents). Thus we had a clear idea how each topic set translation performed in the query system without any idea of the quality of the translations. We also had as a baseline MAPs from monolingual runs from Kettunen (2008). Monolingual and CLIR runs were

TABLE 1: MT programs and their origins

Lang. pair	**MT program**	**Source**	**N.B.**
En → De	Promt Reverso	http://translation2.paralink.com/	
	Google Translate Beta	http://translate.google.com/translate_t	
	Babelfish	http://babelfish.yahoo.com/translate_url	Systran's MT engine
	Translate It!	Timeworks Inc.	A DOS program from 1993
	LEC Translate2Go	http://www.lec.com/t2g_text.asp	
	IBM WebSphere	http://www-01.ibm.com/software/pervasive/tech/demos/translation.shtml	
	SDL Enterprise Translation Server	http://www.freetranslation.com/	
	Systran	http://www.systran.co.uk/	
	InterTran	http://intertran.tranexp.com/Translate/result.shtml	
	Translated	http://free.translated.net/	
	Hypertrans	http://www.dagostini.it/hypertrans/index.php	Patent translator
	MZ Win Translator	http://www.mz-translator.de/	

submitted with the keywords in their plain forms, only stop-words were omitted from the queries.

3 Results

For better understanding of the translation quality of MT programs we evaluated the translation results of different MT systems with one of the latest machine translation evaluation metrics, METEOR 0.6 (Lavie and Agarwal 2008; Banerjee and Lavie, 2005). METEOR is based on a BLEU (Papineni et al., 2002) 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 (Lavie et al., 2004). 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." (Lavie and Agarwal, 2007). METEOR has been shown to outperform commonly used metrics BLEU and NIST in terms of correlations with human judgements of translation quality (Lavie et al., 2004).

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 is a segment to be evaluated, and an overall figure for all the topic translations is given. Table 2 shows the results of METEOR's evaluations for all the English → German title and description MT outputs in their raw form. Table 3 shows results for title translation evaluations.

The meanings of the metrics in Tables 2 and 3 are as follows:

- *Overall system score* gives a combined figure for the result. It is computed as follows (Lavie and Agarwal, 2005): 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.

[1] If this methodology were to be used e.g. with 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 2: Results of METEOR translation evaluation for German TD topics

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Overall system score	0.32	0.26	0.24	0.19	0.30	0.27	0.26	0.24	0.07	0.28	0.22	0.24
Precision	0.60	0.52	0.51	0.46	0.56	0.54	0.56	0.49	0.32	0.53	0.49	0.50
Recall	0.62	0.56	0.53	0.49	0.58	0.56	0.54	0.51	0.33	0.57	0.51	0.52
Fmean	0.62	0.56	0.53	0.49	0.58	0.56	0.54	0.51	0.33	0.56	0.51	0.52
Penalty	0.49	0.54	0.54	0.60	0.48	0.52	0.52	0.53	0.77	0.51	0.56	0.54

Table legend: 1 = Google Translate Beta, 2 = Babelfish, 3 = Promt Reverso, 4 = Translate It!, 5 = Systran, 6 = LEC Translate2Go, 7= IBM WebSphere, 8= SDL Enterprise Translation Service, 9 = InterTran, 10 = Translated, 11 = Hypertrans 12 = MZ Win Translator

Table 3: Results of METEOR translation evaluation for German T topics

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Overall system score	0.29	0.33	0.20	0.22	0.26	0.29	0.27	0.29	0.13	0.28	0.27	0.26
Precision	0.62	0.63	0.57	0.52	0.55	0.60	0.64	0.60	0.44	0.58	0.61	0.59
Recall	0.66	0.66	0.59	0.57	0.61	0.63	0.61	0.63	0.47	0.61	0.61	0.59
Fmean	0.65	0.66	0.59	0.56	0.61	0.63	0.62	0.63	0.47	0.61	0.61	0.59
Penalty	0.56	0.50	0.65	0.61	0.57	0.54	0.56	0.53	0.73	0.55	0.55	0.57

- *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 (2005) 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 TD queries collected in Table 4, we notice that MAPs of the long query runs are in the order 1> 5 > 2 > 10 > 7 > 12 > 6 > 3 >11 > 8 > 4 >9. GAP, difference between the best and worse MAP is, 20.9 %, but most of the differences in MAPs are small, standard deviation from the mean being 4.8. Google's MAP is the only outstanding performance and Intertran is clearly the worst performer.

Table 5 gives results of T queries and relates MAPs of different MT systems to MT metrics.

Order of systems for T queries by the MAP is 1 > 10 > 5 > 2 > 3 > 8 > 4 > 7 > 11 > 12 > 6 > 9. High and low ends of the scale are again the same and clearly distinguished, and the GAP is 15.8 %. One more system gains lower MAPs than the mean value when compared to TD queries, and the standard deviation from the mean is 4.1.

Relative orders by MAPs and MT qualities for TD queries are given in Table

Table 4: Mean average precisions of translated plain German TD queries and MT metrics scores.

	MAP of TD queries	**Meteor's score**
Google Translate Beta	39.9	0.32
Systran	31.6	0.30
Babelfish	30.3	0.26
Translated	30.1	0.28
IBM WebSphere	28.8	0.26
MZ Win Translator	28.2	0.24
LEC Translate2Go	27.8	0.27
Promt Reverso	27.5	0.24
Hypertrans	27.0	0.22
SDL Enterprise Translation Server	26.7	0.24
Translate It!	26.1	0.19
InterTran	19.0	0.07
Mean value	**28.6**	**0.24**
Standard deviation	**4.8**	**0.06**
Monolingual baseline	**38.4**	

6 and for T queries in Table 7.

If we study the orders of TD and T queries given by MAPs and MT quality in Tables 6 and 7, we see that TD queries are given a more consistent order by both measures. There are 6 differences in the order of TD queries in Table 6, when there are 11 differences in the order of T queries in Table 7. Five of the MTQ figures are ties in TD queries and 7 in T queries. Changes of list order are grosser with T queries, as seen from the third column figures of both tables.

TD queries have two positive and two negative changes in the list order, if we consider changes that are bigger than 2 positions. LEC Translate2Go gains 3 positions and so does SDL Enterprise Translation Server. LEC Translate2Go has a relatively high recall value, 0.56, when the mean for all systems in TD queries is 0.53, and its penalty is also low, 0.51 (mean being 0.55) as seen in Table 2. SDL Enterprise Translation Server does not get high recall or Fmean (both 0.51), but its penalty is lower than the mean, 0.53. Probably this gives it a boost in the list order.

METEOR's evaluation results of short queries differ more from the MAP order, and there are 11 differences in the orders. Google's MAP, for example, is much better than Babelfish's with T queries, but Babelfish is given a better MTQ score. A closer examination of the figures in Table 2 reveals that Google's penalty score with T queries is much higher than Babelfish's, but they have the same recall and almost the same Fmean. 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

TABLE 5: Mean average precisions of translated plain German T queries and MT metrics scores.

	MAP of T queries	**Meteor's score**
Google Translate Beta	30.1	0.29
Translated	25.8	0.28
Systran	25.7	0.26
Babelfish	24.2	0.33
Promt Reverso	21.4	0.20
SDL Enterprise Translation Server	20.6	0.29
Translate It!	20.5	0.22
IBM WebSphere	20.5	0.27
Hypertrans	20.4	0.27
MZ Win Translator	19.2	0.26
LEC Translate2Go	18.8	0.29
InterTran	14.3	0.13
Mean value	**21.8**	**0.26**
Standard deviation	**4.1**	**0.05**
Monolingual baseline	**28.5**	

TABLE 6: Order of systems by MAPs and MT quality, TD queries. Same order marked with bold.

Order by MAP	**Order by MTQ**	**Change (2nd column relative to 1st)**
Google Translate Beta	**Google Translate Beta**	**0**
Systran	**Systran**	**0**
Babelfish	Translated	+1
Translated	LEC Translate2Go	+3
IBM WebSphere	**IBM WebSphere**	**0**
MZ Win Translator	Babelfish MT progam	-3
LEC Translate2Go	SDL Enterprise Translation Server	+3
Promt Reverso	**Promt Reverso MT program**	**0**
Hypertrans	MZ Win Translator	-3
SDL Enterprise Translation Server	Hypertrans	-1
Translate It!	**Translate It! MT program**	**0**
InterTran	**InterTran**	**0**

Table 7: Order of systems by MAPs and MT quality, T queries. Same order marked with bold.

Order by MAP	**Order by MTQ**	**Change (2nd column relative to 1st)**
Google Translate Beta	Babelfish MT progam	+3
Translated	Google Translate Beta	-1
Systran	LEC Translate2Go	+8
Babelfish	SDL Enterprise Translation Server	+2
Promt Reverso	Translated	-3
SDL Enterprise Translation Server	IBM WebSphere	+2
Translate It!	Hypertrans	+2
IBM WebSphere	Systran	-5
Hypertrans	MZ Win Translator	+1
MZ Win Translator	Translate It! MT program	-3
LEC Translate2Go	Promt Reverso MT program	-6
InterTran	**InterTran**	**0**

is apparent that the difference in the overall system score is due to the differences in the penalty score, as other scores of Google are quite close to Babelfish's. The same explanation seems to hold for all the dips in the MTQ order: Systran falls 5 positions with a relatively good Recall of 0.61, as its penalty is 0.57. Promt Reverso's recall is 0.59 and penalty 0.65, and it is 6 positions lower in the MTQ order than in MAP order. The same holds for Translate It! Only Translated's dip of positions seems not be caused by higher penalty: its penalty is only 0.55 and recall quite high, 0.61. The biggest gains in MTQ order seem also to follow this pattern. LEC Translate2Go is quite low in MAP order, but gains 8 positions in MTQ order with recall of 0.63 and penalty of 0.54.

Word order of translations is relevant from a translation point of view but it does not affect IR results (Kraaij, 2001), 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 seems also able to indicate the best title translations and worst title translations, although the order of evaluation results differed from the retrieval result order due to metric's inner logic.

Correlation coefficient for MAPs of TD queries and METEOR overall scores is high: **0.88.** Correlation coefficient for T query MAPs and METEOR scores is lower than for TD queries**,** but still clear**, 0.59**. Both correlations were statistically highly significant, when pairwise t-test was used. If we calculate the correlation coefficient disregarding effect of word order penalty (this means in practice, that

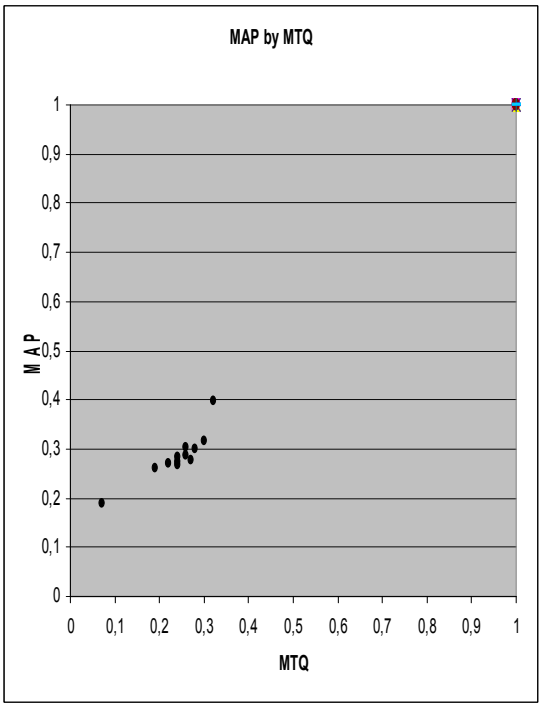

FIGURE 1: Correlation of TD query MAPs and MTQ scores

the Fmean score is the overall score for the translation quality), we get correlation coefficient of **0.85** for TD queries and **0.68** for T queries. Word order penalties seem to thus affect results of T query evaluation more.

The correlations and their difference with TD and T query MAPs and MTQ scores can be seen more clearly from Figures 1 and 2.

Now we can turn to the "predictive capability" of MTQ results. Tables 4 and 5 show, that there are six MT systems whose MTQ score is over the mean value (Babelfish, Google, IBM Websphere, LEC Translate2Go, Systran and Translated) with TD queries and seven with T queries (Babelfish, Google, Hypertrans, IBM Websphere, LEC Translate2Go, SDL Enterprise Translation Server and Translated). Five out of the MTQ score picked six systems in TD queries yield MAPs that are over the mean value, the only exception being LEC Translate2Go. Three MTQ score picked systems in T queries yield MAPs over the mean, Google, Babelfish and Translated, other four perform under the mean. Systran gets MAP that is over the mean, but its MTQ value is the mean. Thus it seems that predictive power of the MTQ score is better with TD queries and more fluctuating with T queries.

One further aspect that should be taken into account with respect to T queries is the length of the topics. A common belief in CLIR research is, that as queries are many times short and not even full sentences, their translation with MT programs

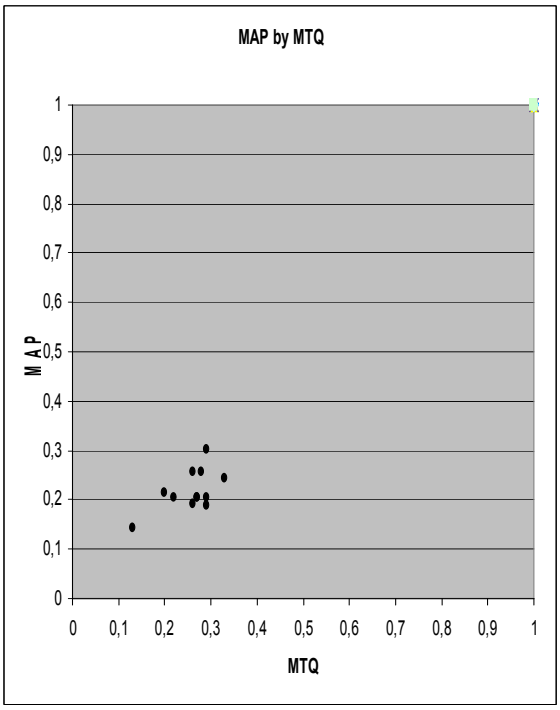

FIGURE 2: Correlation of T query MAPs and MTQ scores

is difficult or problematic in some way (cf. e.g. Kishida, 2005). As we translated T parts of topics and TDs separately, we noticed that translations of Ts and beginnings of TDs are most of the times the same. Thus it seems that the common belief does not hold in an IR laboratory type of evaluation setting of short queries. To study this, we compared beginnings of translated **60** TDs to translated Ts with Ultraedit's Ultra Compare Professional, a character level comparison software. Table 8 lists results of comparisons.

Results show that 8 systems out of 12 translate title parts of the topics in the same way as the beginnings of TDs regardless of the following context (or its absence). Two of the systems that have translation differences (Promt Reverso, Hypertrans) translate the beginnings most of the time identically (44 and 53 identical translations), and only two systems (Babelfish and Systran, that use the same MT engine) have about half of the translations differing (26 and 31 identical translations).

Thus the length of the queries with respect to MT's translation results is not an issue here, because translations of T queries are almost always the same as the beginnings of TD queries even when T parts are translated separately. Lengths of different translations correlate also. With TD translations the correlation is 0.86 and with T queries 0.61, when translations of six MT systems were correlated against translations of other six systems. There is a slight decrease of MTQ mean

TABLE 8: Comparison of T translations to beginnings of TD translations

	TD translation's beginnings compared to T translations
Google Translate Beta	no differences
Systran	Quite many differences, but many of them are differences in word order, choice of pre- and postpositions. Also differences in vocabulary, e.g. *Risiken mit Handys. /Gefahren mit beweglichen Telefonen // Dayton Friedensvertrag./ Dayton Friedensabkommen.* **26** identical translations.
Babelfish	Many differences, but many of them are differences in word order, choice of pre- and postpositions. Also differences in vocabulary, e.g. *Aufkommen des CD Burner / Aufkommen des CD Brenners // Hubble und schwarze Bohrungen. / Hubble und schwarze Löcher.* **31** identical translations.
Translated	no differences
IBM WebSphere	no differences
MZ Win Translator	no differences
LEC Translate2Go	no differences
Promt Reverso	Minor differences. Some vocabulary differences, e.g. *Holländische Fotos von Srebrenica / Niederländisch Fotos von Srebrenica.* **44** identical translations.
Hypertrans	Minor differences in vocabulary, e.g. *Französische allgemeine und Balkan-Sicherheit Zone / Französischer General und Balkan-Sicherheit Zone // Rücktritt von NATO-Sekretärin allgemein / Rücktritt von borner Sekretärin General.* **53** identical translations.
SDL Enterprise Translation Server	no differences
Translate It!	no differences
InterTran	no differences

score in TD query translations in comparison to T query translations, 0.24 vs. 0.26, as was seen in Tables 4 and 5 (using Fmean as the figure we get 0.53 vs. 0.60). This is most obviously caused by the fact, that the mean length of T queries is 3.7 words and the mean length of TD queries 18.8 words. With TD queries MT systems have more options to translate the queries differently from the CLEF human translation, which is seen as a lower MTQ mean score. Thus there are two opposing tendencies that affect the results: the length of queries affects conversely MTQ scores and MAPs: short query translations get better MTQ scores and lower MAPs and vice versa. The relation of achieved MAP and query length is a known issue in IR, but the relation of MT quality and query length is opposed to the common belief in CLIR literature, where short queries are considered a harder translation task for MT programs.

4 Discussion and conclusions

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. Word order of translations, for example, does not affect IR results (Kraaij, 2001). We evaluated twelve English German translations 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. Also the worst MT system was clearly indicated. The MTQ score was able to pick five MT systems out of six that achieved best MAPs with TD queries. 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. With T queries MTQ scores picked three systems achieving MAPs that were over the mean value, but the scores also picked four systems that performed under the mean MAP. The worst MT system was also clearly distinguished by the MTQ score. T queries got thus more fluctuating scores by METEOR than by MAP.

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 length of the queries affects results. 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. Based on the findings of the paper 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. With longer laboratory type

queries the task is easier, and with short queries the results are more varying. 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, if the metrics uses that, because is not relevant in this use. Perhaps also some other fine-tuning could be needed for MT metrics in this specific use.

In this study we were able to use 12 MT systems to produce translations. For the most common language pairs, such as English, German, French and Spanish, this plentitude of available MT systems is a reality, but for other language pairs fewer MT systems are usually available. However, in Kettunen (2009) we showed with four MT systems same types of correlations between MAPs and MTQ scores. We also showed that MAPs and MTQ scores of two other systems, BLEU and NIST, correlated. Thus we believe that the approach is also useful with other MT metrics and when the language pair has available fewer MT systems.

References

Banerjee, S. & Lavie, A. (2005). 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).

Church, K. W. & Hovy, E.H. (1993). Good application for crummy machine translation. Machine Translation, 8, 239–258.

Kettunen, K. (2008). MT-based query translation CLIR meets Frequent Case Generation. Submitted.

Kettunen, K. (2009). Choosing the best MT programs for CLIR purposes - can MT metrics be helpful? In M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, 706–712.

Kishida, K. (2005). Technical Issues of Cross-Language Information Retrieval: A Review. Information Processing & Management, 41, 433–455.

Kishida, K. (2008). Prediction of performance of cross-language information retrieval system using automatic evaluation of translation. Library & Information Science Research, 30, 138–144.

Kraaij, W. (2001). TNO at CLEF-2001: Comparing Translation Resources. In Working Notes for the CLEF 2001 Workshop. Retrieved September 15, 2008, from http://www.ercim.org/publication/ws-proceedings/CLEF2/kraaij.pdf.

Lavie, A. & Agarwal, A. (2007). 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).

Lavie, A. & Agarwal, A. (2008) The METEOR Automatic Machine Translation Evaluation System, http://www.cs.cmu.edu/~alavie/METEOR/

Lavie, A., Sagae, K., & Jayarman, S. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), Washington, DC, (pp. 134–143).

McNamee, P. & Mayfield, J. (2002). Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In Proceedings of Sigir'02, Tampere, Finland, (pp. 159–166).
Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J: (2002). BLEU: a method for automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), (pp. 311–318).
Zhu, J. & Wang, H. (2006). 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).