Evaluation of the domain adaptation of MT systems in ACCURAT

Gregor Thurmair
Linguatec,
Gottfried-Keller-Str. 12, 81375 Munich, Germany
gregor.thurmair@gmx.de

Abstract[1]

The contribution reports on an evaluation of efforts to improve MT quality by domain adaptation, for both rule-based and statistical MT, as done in the ACCURAT project (Skadiņa et al. 2012). Comparative evaluation shows an increase of about 5% for both MT paradigms after system adaptation; absolute evaluation shows an increase in adequacy and fluency for SMT. While the RMT solution is superior in quality in both comparative and absolute evaluation, the gain by domain adaptation is higher for the SMT paradigm.

1 Introduction

The objective of this contribution is to evaluate improvements achieved by adapting Machine Translation systems to narrow domains, using data from comparable corpora.
Language direction chosen was German to English; the automotive domain, subdomain of transmission / gearbox technology, was selected as an example for a narrow domain. In order to assess the effect of domain adaptation on MT systems with different architecture, both a data driven (SMT) and a knowledge-driven (RMT) system were evaluated.

2 Evaluation Objects: MT systems adapted to narrow domains

The evaluation object are two versions of an MT system: A baseline version, *without* domain tuning, and an adapted version, *with* domain tuning. Their comparison shows to which extent the domain adaptation can improve MT quality.
The evaluation objects were created as follows:

For the **baseline** systems, on the RMT side, an out-of-the-box system of Linguatec's 'Personal Translator' PT (V.14) was used, which is a rule-based MT system, based on the IBM slot-filler grammar technology (Aleksić & Thurmair 2011) and a bilingual lexicon of about 200K transfers. On the SMT side, a baseline Moses system was trained with standard parallel data (Europarl, JRC etc.), plus some initial comparable corpus data as collected in the first phase of ACCURAT.

For the **adaptation** of the baseline systems, data were collected from the automotive domain. These data were obtained by crawling sites of automotive companies being active in the transmission field (like ZF, BASF, Volkswagen and others), using the focused crawler described in (Papavassiliou et al. 2013). They were then aligned and cleaned manually. Some sentence pairs were set aside for testing, the rest was given to the two systems as development and test sets. The resulting narrow-domain automotive corpus has about 42.000 sentences for German-to-English.

For the SMT system, domain adaptation was done by adding these sentences to the training and development sets, and building a new SMT system.

In case of rule-based technology, domain adaptation involves terminology creation, as the main means of adaptation. The following steps were taken: (1) extraction of the phrase table from the just described domain-adap6ted SMT system; (2) extraction of bilingual terminology candidates from this phrase table, resulting in a list of about 25.000 term candidates; (3) preparation of these candidates for dictionary import; creation of linguistic annotations, removal of already existing entries etc.; the final list of imported entries was about 7100 entries; (4) crea-

[1] This research was funded in the context of the FP7-ICT project ACCURAT (248347),

tion of a special 'automotive' user dictionary, to be used additionally for automotive translations. This procedure is described in detail in (Thurmair & Aleksić 2012).
Result of these efforts were four test systems, for German-to-English, and tuned for automotive domain with the same adaptation data:
SMT-base: Moses with just baseline data
SMT-adapted: Moses baseline plus in-domain data
RMT-base: PT-baseline out-of-the-box
RMT-adapted: PT with an automotive dictionary.

3 Evaluation Data

In total about 1500 sentences were taken from the collected strongly comparable automotive corpora for tests, with one reference translation each. The sentences represent 'real-life' data; they were not cleaned or corrected..

4 Evaluation methodology

4.1 General options

Several methods can be applied for the evaluation of MT results, cf. Figure 1.

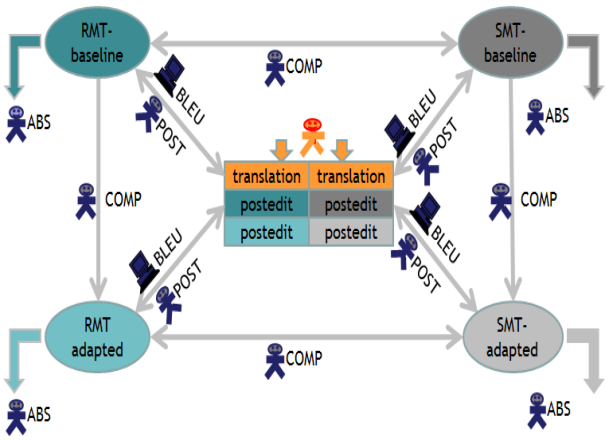

Fig. 1: Evaluation graph

1. Automatic comparison (called BLEU in Fig. 1) is the predominant paradigm in SMT. BLEU (Papineni et al. 2002) and/or NIST (NIST 2002) scores can be computed for different versions of MT system output. Because of their known shortcomings (Callison-Burch et al. 2009) evaluations ask for human judgment in addition.
2. Comparative evaluation (called COMP in Fig. 1) compares two systems, or two versions of the same system. It asks whether or not one translation is better / equal / worse than the other.

While this approach can find which of two systems has an overall better score, it cannot answer the question what the real quality of the two systems is: 'Equal' can mean that both sentences are perfect, but also that both are unusable.
3. Absolute evaluation (called ABS in Fig. 1) therefore is required to determine the quality of a given translation. It looks at one translation of a sentence at a time, and determines its accuracy and fluency on a n-point scale.
4. Postediting evaluation (called POST in Fig. 1) reflects the task-oriented aspect of evaluation (Popescu-Belis 2008). It measures the distance of an MT output to a human (MT-postedited) output, either in terms of time, or of the keystrokes needed to produce a corrected translation from a raw translation (Tatsumi 2009; HTER: Snover et al. 2009).

Postediting evaluation adds reference translations to the evaluation process.

The evaluation graph as shown in Fig. 1 combines these evaluation methods, avoids biased results as produced by a single method, and gives a complete picture of the evaluation efforts.

4.2 Evaluation in ACCURAT

In the ACCURAT narrow domain task, the following evaluation methods were used:
Automatic evaluation of the four systems (SMT and RMT, baseline and adapted) using BLEU.
Comparative evaluation of the pairs SMT-baseline vs. SMT-adapted, and RMT-baseline vs. RMT-adapted; this produces the core information how much the systems can improve.
Absolute evaluation of the systems SMT-adapted and RMT-adapted, to gain insight into translation quality, and consequently the acceptance of such systems for real-world use.

Other forms of evaluation were not included, esp. postediting evaluation was done in other tasks in the ACCURAT project (cf. Skadiņš et al. 2011). But to have a complete picture, other ABS and COMP directions were evaluated, but with 1 tester only.

For the evaluation, a special tool was created called 'Sisyphus II', to be used offline by freelancers, randomly proposing evaluation data, and creating an XML file for later evaluation.

5 Evaluation Results

5.1 Automatic Evaluation

The automatic evaluation for the four test systems was done using BLEU scores. The results are shown in Table 1.
For both system types there is an increase in BLEU; more moderate for the RMT than for the SMT system. Also, the SMT system performs

better in this evaluation method. However it is known that BLEU is biased towards SMT systems (Hamon et al. 2006, Culy & Riehemann 2003).

	SMT	RMT
baseline	17.36	16.08
adapted	22.21	17.51
improvement	4.85	1.43

Table 1: BLEU scores for SMT and RMT

5.2 Comparative Evaluation

Three testers were used, all of them good speakers of English with translation background. They inspected randomly selected subsets of the 1500 test sentences. Results are given in Tab. 2.

	total inspected	base better	both equal	adapted better	improvement
Total SMT	2060	447	1061	552	5.10%
Total RMT	2505	158	2072	275	4.67%

Table 2: Comparative Evaluation: baseline vs. adapted, for SMT and RMT[2]

Both types of systems show an improvement of about 5% after domain adaptation. It is a bit more for the SMT than for the RMT, due to a strong RMT baseline system.

The result is consistent among the testers: All of them see a higher improvement for the SMT than for the RMT.

It may be worthwhile noticing that in the RMT evaluation, a large proportion of the test sentences (nearly 60%) came out identical in both versions. In the SMT system, nearly no sentence came out unchanged; this fact increases the postediting effort for consecutive versions of SMT output.

In a sideline evaluation, a comparison was made between the RMT and SMT systems, for both baseline and adaptations, cf. Tab. 3.

	total inspected	SMT better	both equal	RMT better	in %
baseline	501	47	170	284	47.3%
adapted	489	38	203	260	44.3%

Table 3: Comparative Evaluation SMT vs. RMT, for baseline and adapted

The result shows that the RMT quality is considered significantly better than the SMT quality. The main reason for this seems to be that the SMT German-English frequently eliminates verbs in sentences, which makes the output much less understandable.

It should be noted, however, that the distance between the system types is smaller in the adapted than in the baseline versions (by 3%).

5.3 Absolute Evaluation

Absolute evaluation assesses how usable the resulting translation would be after the system was adapted. A total of 1100 sentences, randomly selected from the 1500 sentence test base, were inspected by three testers for adequacy and fluency. Table 4 gives the result.

	inspected	1: fully	2: mostly	3: partially	4: none	average	% of 1+2
		SMT adapted					
adequacy	1103	200	204	517	182	2.62	36,63
fluency	1103	300	285	362	156	2.34	53,04
		RMT adapted					
adequacy	1103	300	285	362	156	2.02	53,04
fluency	1103	300	285	362	156	1.80	53,04

Table 4: Absolute evaluation for SMT-adapted and RMT-adapted systems[3]

It can be seen that testers evaluate the SMT somewhat between 'mostly' and 'partially' fluent / comprehensible, and the RMT close to 'mostly' fluent / comprehensible. If the percentage of level 1/2 evaluations is taken, both adequacy and fluency rates are significantly higher for RMT output. All testers agree in this evaluation, with similar average results.

It could be worthwhile to mention that the opinion often heard that the SMT produces more fluent output that the RMT cannot be corroborated with the evaluation data here: The RMT output is clearly considered to be more fluent than the SMT output (1.80 vs. 2.34).

As far as the interrater agreement is concerned, the test setup made it difficult to compute it: All testers used the same test set but tested only a random subset of it. So there are only few data points common to all testers (only 20 in many cases). For those, only weak agreement could be found (with values below 0.4 in Cohen's kappa). However, all testers show consistent behaviour in the evaluation, and came to similar conclusions overall, as has been explained above.

[2] Computed as: (#-better MINUS #-worse) DIV #-sentences

[3] Computed as: (SUM (#-sentences TIMES rank)) DIV #-sentences. Lower scores are better.

6 Conclusion

Figure 5 gives all evaluation results. All evaluation methods indicate an improvement of the adapted versions over the baseline versions.

Automatic evaluation: For SMT, the BLEU score increases from 17.36 to 22.21; for RMT, it increases from 16.08 to 17.51.

Comparative evaluation: For SMT, an improvement of 5.1% was found; for RMT, and improvement of 4.67% was found.

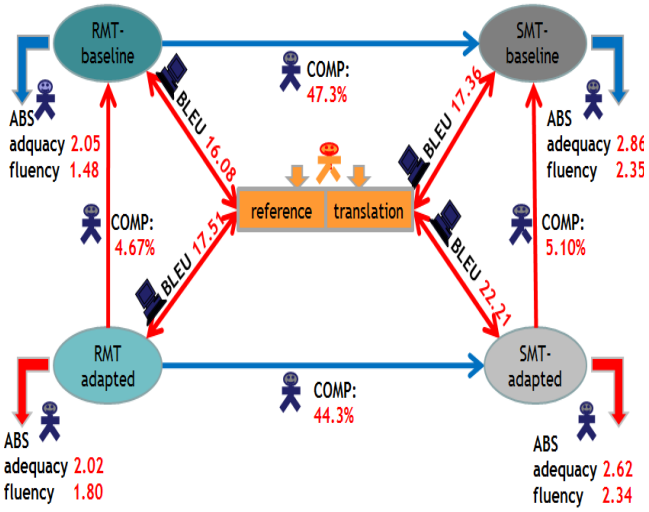

Fig. 5: Evaluation graph for ACCURAT task

Absolute evaluation: For SMT, adequacy improved from 2.86 to 2.62, fluency slightly from 2.35 to 2.34; for RMT, adequacy improved from 2.05 to 2.02, fluency decreased from 1.48 to 1.8.

The improvement is more significant for the SMT system than for the RMT; this may be due to the fact that the RMT baseline system was stronger than the SMT baseline.

For SMT improvement, (Pecina et al. 2012) report improvements between 8.6 and 16.8 BLEU (relative) for domain adaptation; results here are in line with these findings.

Comparing the evaluation methods, the findings corroborate statements (cf. Hamon et al. 2006) that the 'human-based' methods (COMP and ABS) differ from the automatic ones (BLEU) if different types of MT systems are to be compared.

Overall, the 'human-based' evaluation methods (COMP, ABS) have shown that a trained RMT system still outperforms a trained SMT system; however the SMT system profits more from adaptation.

References

Aleksić, V., Thurmair, Gr., 2011: Personal Translator at WMT 2011. Proc. WMT Edinburgh, UK.

Callison-Burch, Chr., Koehn, Ph., Monz, Ch., Schroeder, J., 2009: Findings of the 2009 Workshop on Statistical Machine Translation. Proc 4th Workshop on SMT, Athens

Culy, Chr., Riehemann, S.,2003: The Limits of N-Gram Translation Evaluation Metrics. Proc. MT Summit New Orleans

Hamon, O., Popescu-Belis, A., Choukri, K., Dabbadie, M., Hartley, A., W. Mustafa El Hadi, W., Rajman, M., and Timimi, I., 2006. CESTA: First Conclusions of the Technolangue MT Evaluation Campaign. Proc. LREC Genova, Italy.

NIST (2002). Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrences Statistics (http://www.nist.gov/speech/tests/mt/)

Papavassiliou, V., Prokopidis, P., Thurmair, Gr., 2013: A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. Proc. 6th Workshop BUCC, Sofia, Bulgaria

Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002: BLEU: a Method for Automatic Evaluation of Machine Translation. Proc. ACL, Philadelphia

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., van Genabith, J., 2012: Domain Adaptation of Statistical machine Translation using Web-Crawled Resources: A Case Study // Proceedings of the EAMT 2012, Trento, Italy.

Popescu-Belis, A., 2008: Reference-based vs. task-based evaluation of human language technology. Proc. LREC

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M.L., Pinnis, M., 2012: Collecting and Using Comparable Corpora for Statistical Machine Translation. Proc. 8th LREC Istanbul

Skadiņš, R., Puriņš, M., Skadiņa, I., Vasiļjevs, A., 2011: Evaluation of SMT in localization to under-resourced inflected language. Proc. EAMT Leuven

Snover, M., Madnani, N., Dorr, B., Schwartz, R., 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In Proc. of WMT09

Tatsumi, M., 2009: Correlation Between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. Proc. MT Summit XII Ottawa

Thurmair, Gr., Aleksić, V., 2012: Creating Term and Lexicon Entries from Phrase Tables. Proc. EAMT 2012 Trento, Italy.