MT goes farming: Comparing two machine translation approaches on a new domain

Per Weijnitz, Eva Forsbom, Ebba Gustavii, Eva Pettersson, Jörg Tiedemann

Department of Linguistics and Philology
Uppsala University, Uppsala/Sweden
{perweij,evafo,ebbag,evapet,joerg}@stp.ling.uu.se

Abstract

In the paper we present detailed analyses of two machine translation systems when applied to documents of a previously unseen domain: agricultural texts from the European Union. The two systems compared are a statistical machine translation (SMT) system using the freely available ISI ReWrite Decoder (Germann, 2003a), and the rule-based machine translation system MATS (Sågvall Hein et al., 2002). For the purpose of comparison we use a sentence-aligned Swedish-English corpus of approximately 75,000 words per language, where 90% are used for training and 10% are used for evaluation. In the paper we discuss the outcome of automatic evaluation and the results of our manual quality assessment.

1. Introduction

Knowledge-rich rule-based models and knowledge-poor statistical models are two different approaches to machine translation (MT). In this study, we compare two existing MT systems, one from each paradigm. The aim is to find out which of the two systems should be preferred for achieving editing quality in a new domain, using limited resources. The two systems compared are a statistical machine translation (SMT) system using the freely available ISI ReWrite Decoder (Germann, 2003a), and the rule-based machine translation (RBMT) system MATS. (Sågvall Hein et al., 2002). In the paper we present detailed analyses of both systems when applied to agricultural texts from the European Union. We will discuss the outcome of automatic evaluation and manual quality assessment[1].

2. SMT and SMT tools

In SMT, translation is modeled as the transmission of a sentence t in language T through a noisy channel that changes the signal into a sentence s in language S (Brown et al., 1993). The task in SMT is to find the string t', which is most likely the original string that has been transmitted when observing s. The fundamental search problem in SMT is defined as follows: $t' = argmax_t P(s|t)P(t)$

Two models have to be found, the translation model $P(S|T)$ and the language model $P(T)$ for the target language T. In our experiments, we use the freely available toolbox GIZA++, version 2 (09/30/2003), (Och and Ney, 2000) for training the translation model and the CMU-Cambridge Statistical Language Modeling Toolkit v2 (Clarkson and Rosenfeld, 1997) for creating a language model for the target language. We use the standard settings for both tools, i.e. IBM model 1 to 4 for the translation model and a trigram model for the target language. Alignment models depending on word classes apply classes that have been found automatically with the *mkcls* tool (Och, 1999).

For the actual translation we used the publicly available ISI ReWrite Decoder Release 1.0.0a (Germann, 2003a). This software implements several optimised strategies for fast decoding of statistical translation models. It uses IBM model 4 and expects language models created by the Language Modeling Toolkit that we used in our experiments. Several parameters can be given to the decoder in order to test different algorithms and to produce different amounts of debugging output. It implements five decoding algorithms, three variants of gloss maximisation and two versions of greedy decoding algorithms (fast and thorough). Furthermore, parameters can be specified to optimise search space options. More information can be found in the decoder manual (Germann, 2003b).

3. RBMT and MATS

In rule-based MT systems, translation is based on formalised linguistic knowledge, represented in dictionaries and grammars. MATS (Sågvall Hein et al., 2002) is a research prototype in the traditional transfer paradigm, with separate modules for source language analysis, transfer, and target language generation. The analysis module uses the procedural Uppsala Chart Parser. Transfer and generation are handled by MULTRA, a unification-based translation engine. MATS has been developed primarily for translating automotive service literature. Its grammar has been trained and evaluated using a parallel corpus of approximately 50,000 tokens (the *MATS corpus*).

A general problem with rule-based systems is robustness. Translation usually fails if the input is not covered by the grammars or dictionaries. MATS has recently been extended with mechanisms that make use of incomplete parses and structures not covered by transfer or generation grammars. This back-off technique improves robustness of the system and makes it easier to adapt to new domains.

4. Adaptation to the new domain

The parallel corpus representing the new domain comprises agricultural reports, specifications and circulars produced within the European Union (the *AGRI corpus*). The documents were provided by the European Commission

Translation Service (SDT) within the project *Extension of EC Systran to Danish and Swedish into English, Commission contract SDT/MT2003-1*. The language exhibits features typical of official documents i.e. an extensive use of subordinate clauses, contractions and abstract verbal nouns (Cassirer, 1995).

The corpus contains 6732 aligned text segments (mostly one-to-one sentence alignments) with about 71,000 Swedish tokens and 86,000 English tokens (counting punctuations as tokens). Each segment contains about 10.6 Swedish tokens and about 12.8 English tokens on average. As expected, sentences in the agricultural domain are significantly longer compared to the domain we were previously working on (technical manuals) with about 5.9 Swedish tokens per segment and about 7.7 English tokens per segment. We expect this to have a strong impact on the quality of the translations as longer segments are generally harder to translate.

4.1. SMT

Adapting statistical machine translation to a new domain is a matter of training probabilistic models on new material. For our purposes, this task included the installation and preparation of software and tools that are needed for such a training step. Recently, such tools became publicly available as described in section 2. These tools are optimised for working together especially on a standard GNU/Linux platform. They include documentation for building a working SMT application from training to translation. Therefore, installing and preparing the system on our GNU/Linux machines did not cause larger difficulties.

Training SMT models is easy using the standard settings of training modules described in the documentation. However, several pre-processing steps have to be performed to adjust input formats and system settings. A number of simple scripts have been implemented in order to make the training procedure easier. Once this had been done, new models could be created with a single command for different variants of training data. Optimising training and decoding is by far more difficult. There are many options for both the training software (translation model and language model) and the decoding software.

Altogether we estimate the time for installing and training the system (including tests with different parameter settings and the implementation of some helper tools) to about 2 weeks.

4.2. RBMT

Whereas the adaptation of a statistical machine translation system is an automatic procedure, adapting a rule-based machine translation system generally involves time-consuming manual work. For an optimal result, both the dictionaries and the grammars need to be tuned to the new domain. Due to time-constraints, only the dictionaries have been adapted in our experiment, leaving room for further improvements at a later stage.

The Swedish-English dictionary was compiled and fine-tuned as part of the project *Extension of EC Systran to Danish and Swedish into English*. The dictionaries are corpus-based and defined semi-automatically with the use of word

alignment techniques developed by Tiedemann, 2003). The alignment process included both statistical information and linguistic clues, such as part-of-speech tagging and chunk parsing.

The dictionary work amounts to an effort of approximately four person-months, including planning and development of suitable methods (not including the word alignment software).

5. Evaluation methodology

5.1. Automatic evaluation

For comparing the outcome of the two systems we utilise four automatic evaluation measures that compare the similarity of a candidate translation to a reference translation. We use measures based on both n-gram co-occurrence and edit distance.

BLEU is one of the n-gram measures that has frequently been used in MT evaluations. It is based on the average of matching n-grams between a proposed translation and one or more reference translations, and it seems to correspond well with human judgments on accuracy and fluency (Papineni et al., 2001). An alternative measure is NEVA, which addresses two irregularities in BLEU for segment level evaluation: 1) BLEU does not correctly handle segments shorter than the largest n-gram defined in the measure (usually 4 words). 2) BLEU uses the geometric mean which results in a score of 0 if there is no match in one of the n-gram classes (e.g. trigrams) even if there are matching n-grams otherwise (e.g. among bigrams). NEVA takes care of both cases by checking the segment length and using arithmetic means instead of geometric (Forsbom, 2003).

The other two measures are based on edit distance, indicating the amount of work a post-editor would have to do to correct the mistakes in the automatic translation (counted in terms of deletions d, substitutions s and insertions i). These measures are Word Accuracy (WA) which has been used for MT evaluation (Alshawi et al., 1998), and WAFT (Forsbom, 2003), which takes account of an irregularity in WA that can occur if the candidate and reference translations differ in length.

$$\text{WA} = \left(1 - \frac{d+s+i}{l_r}\right), \text{WAFT} = \left(1 - \frac{d+s+i}{\max(l_r, l_c)}\right)$$

l_r = length of reference
l_c = length of candidate translation

NEVA and WAFT represent the two types of automatic evaluation measures we like to include in our experiments, being the most reliable ones. BLEU and WA are shown for the sake of reference.

5.2. Manual evaluation

The manual evaluation is guided by the SAE J2450 measure (SAE, 2001). SAE J2450 provides a standard for tagging translation errors according to their type and severity, and a method for mapping the error tags to numeric scores (see figure 1). The meta-rules are: 1) When an error is ambiguous, always choose the earliest primary category, and 2) When in doubt, always choose serious over minor.

The measure was established to enable comparable quality assessments, regardless of language and how the translation is created. Apart from the numeric scores, the manual evaluation results in sets of classified errors. These may be used for estimating how hard it would be to correct the errors in the two systems.

Category c	serious $w_{c,s}$	minor $w_{c,m}$
Wrong term (WT)	5	2
Syntactic error (SE)	4	2
Omission (OM)	4	2
Word structure or agreement error (SA)	4	2
Misspelling (SP)	3	1
Punctuation (PE)	2	1
Miscellaneous error (ME)	3	1

$$\text{SAE J2450} = \frac{1}{N} \sum_c \left(s_c \cdot w_{c,s} + m_c \cdot w_{c,m} \right)$$

s_c = number of serious errors in the category c
m_c = number of minor errors in the category c
N = number of words in the source text

Figure 1: SAE J2450: error categories and computation.

6. Experiments

After adapting the systems, we applied the evaluation corpus to both systems. MATS was used with the back-off mechanism. The ISI ReWrite Decoder was run with the built-in decoding strategies and a variety of models trained on data from the AGRI corpus and the MATS corpus. The SMT system was trained and tested on the original text as well as on a lower-case version. A test using additional training data of about one million tokens from the EU-ROPARL corpus (Koehn, 2003) was also carried out.

For each setting we calculated the automatic measures as described above. The n-gram measures are usually used with multiple reference translations. In our experiment only one reference translation was available, causing lower and less confident scores.

The manual evaluation was carried out by four evaluators using the best result achieved by each system according to the automatic measures (without EUROPARL in SMT). Every 5th segment (=20%) of the test corpus was checked by two evaluators using the SAE J2450 guidelines, shifting one evaluator for every segment to make it possible to compute correlations between evaluations and evaluators.

6.1. Results from automatic evaluations

In general the scores are rather low (see figure 2), partly due to the use of one single reference translation. Another source could be a low terminology overlap between the training and the evaluation corpus.

The SMT system performed best with the greedy2 decoding algorithm without converting to lower-case. According to the edit distance measures, WA and WAFT, the model trained on the combined AGRI and MATS corpora performed the best. Surprisingly, using the more than ten times larger corpus taken from the EUROPARL combined
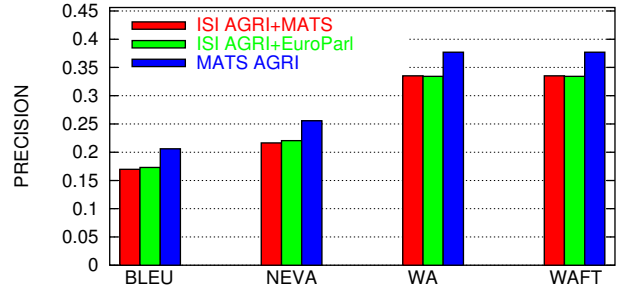

Figure 2: Results from the automatic evaluations.

with AGRI did not noticeably improve the results; only a small increase in terms of NEVA and BLEU scores could be observed. This indicates that SMT cannot be improved simply by adding more training data from another domain.

MATS performed better than the SMT system according to all four evaluation measures.

6.2. Results from manual evaluations

Table 1 summarises the manual evaluations by means of SAE J2450 categories and weighted scores. There is a significant correlation[2] between the two evaluations for the SAE J2450 score measured at segment level (0.7770 for the SMT system and 0.7054 for the RBMT system). Assessments from individual evaluators are also significantly correlated (between 0.6034 and 0.9687, one combination: 0.3721).

RBMT	evaluation 1			evaluation 2		
c	s_c	m_c	score	s_c	m_c	score
WT	365	0	1825	361	0	1805
SE	34	0	136	33	0	132
OM	6	6	36	2	8	24
SA	11	70	184	9	75	186
SP	0	2	2	0	5	5
PE	0	1	1	2	1	5
ME	6	10	28	4	7	19
SAE J2450	overall score		0.9559	overall score		0.9404
SMT	evaluation 1			evaluation 2		
c	s_c	m_c	score	s_c	m_c	score
WT	496	0	2480	512	0	2560
SE	35	0	140	26	0	104
OM	40	9	178	29	14	144
SA	7	45	118	9	41	118
SP	1	0	3	0	1	1
PE	1	2	4	6	2	14
ME	23	8	77	29	4	91
SAE J2450	overall score		1.2706	overall score		1.2853

Table 1: Manual evaluation: SAE J2450 scores.

Wrong term is by far the most commonly attested error type in the output of both systems. Most of these errors involve untranslated words, i.e. items unknown to the systems. SAE J2450 does not distinguish between unknown and wrong terms. Adding a new category to SAE J4520

---

[2]Correlation is measured in terms of linear correlation coefficients.

would make the measure more suitable for MT evaluation. Untranslated words seem to be a larger problem for the SMT system than for MATS. More surprisingly, MATS exposes a greater amount of agreement and word structure errors. For both systems, this error type typically involves noun and verb inflection. MATS mainly encounters these problems when the noun is ambiguous in number or when the subject is not located. For the SMT system the distribution of these errors is less predictable.

The most striking difference between the systems is the amount of omissions; the SMT system exposes more than four times as many instances of omission as MATS.

In general, the cause of the errors produced by MATS could easily be traced and explained in linguistic terms. The SMT system on the other hand, often produces less predictable errors, such as *vegetables* as a translation of *verksamhetsprogram* [operational funds] and *341* as a translation of *uppgick* [amounted]. There is no obvious way to trace the cause of these errors to parameters in the translation or language model.

6.3. Discussion

The rule-based MT system MATS achieved better scores in both the automatic and the manual evaluation. However, it required about eight times as much time to adjust this system as compared to the statistical approach.

Table 2 summarises the correlations between the various manual and automatic evaluations of segments. There is a significant correlation between all types of evaluations,[3] but BLEU has the lowest correlations with all other evaluations.

RBMT	J2450(2)	BLEU	NEVA	WA	WAFT
J2450(1)	0.7054	-0.1788	-0.5039	-0.4518	-0.4727
J2450(2)		-0.1628	-0.5026	-0.4921	-0.5096
BLEU			0.2974	0.3236	0.3221
NEVA				0.9379	0.9409
WA					0.9904
SMT	J2450(2)	BLEU	NEVA	WA	WAFT
J2450(1)	0.7770	-0.2734	-0.6067	-0.6209	-0.6098
J2450(2)		-0.2321	-0.6216	-0.6366	-0.6404
BLEU			0.3842	0.3923	0.3811
NEVA				0.9322	0.9385
WA					0.9738

Table 2: Linear correlation between evaluations.

Automatic and manual evaluation seem to be correlated according to our experiments. Hence, automatic measures seem to be a good approximation of the overall translation quality. However, manual evaluations are still necessary to identify specific weaknesses of existing translation systems.

7. Summary and conclusions

Building a machine translation system requires time and resources. Our experiments show that both the statistical and the rule-based system produce unsatisfactory results

---

[3]Significant for J2450 and BLEU in RBMT at $p < 0.05$ with a standard t-test. All others at $p < 0.01$.

when built in a short period of time and with limited resources. The rule-based system MATS achieved better results but there is much room for improvements in both systems and a lot of common problems to be solved. MATS can be improved by adjusting the grammars and extending the dictionaries. The statistical tools require larger amounts of domain-specific training data for a better coverage and a higher translation quality.

8. References

Alshawi, H., S. Bangalore, and S. Douglas, 1998. Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the 36th ACL*. Montreal, Canada.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cassirer, P., 1995. *Stilistik & Stilanalys*. Bokförlaget Natur och Kultur.

Clarkson, P.R. and R. Rosenfeld, 1997. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings of ESCA Eurospeech*.

Forsbom, E., 2003. Training a super model look-alike: Featuring edit distance, n-gram occurrence, and one reference translation. In *Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation, in conjunction with MT SUMMIT IX*.

Germann, U., 2003a. Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of HLT-NAACL*. Edmonton, AB, Canada.

Germann, U., 2003b. *ISI ReWrite Decoder User's Manual, Version 1.0.0a*.

Koehn, P., 2003. Europarl: A multilingual corpus for evaluation of machine translation. `http://www.isi.edu/~koehn/europarl/`.

Och, F. J., 1999. An efficient method for determining bilingual word classes. In *Proceedings of the 9th EACL*.

Och, F. J. and H. Ney, 2000. Improved statistical alignment models. In *Proceedings of the 38th ACL)*. Hongkong, China.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu, 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176, IBM Research Division, T. J. Watson Research Center.

SAE, 2001. Surface vehicle recommended practice: SAE J2450. Available at `http://www.lisa.org/useful/2001/J2450Practice.pdf`.

Sågvall Hein, A., E. Forsbom, J. Tiedemann, P. Weijnitz, I. Almqvist, L.-J. Olsson, and S. Thaning, 2002. Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings of the 2nd LREC*, volume V. Las Palmas de Gran Canaria, Spain.

Tiedemann, J., 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Department of Linguistics, Uppsala University, Uppsala/Sweden. Acta Universitates Upsaliensis - Studia Linguistica Upsaliensia (1).