A Challenge Set for English-Swedish Machine Translation

Lars Ahrenberg

Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract

This paper presents a project aimed at creating a challenge set for machine translation from English to Swedish. A challenge set is a test suite where sentences or short text snippets with their translations have been selected for purposes of evaluation. The current version contains 202 cases covering various translation problems in the direction from English to Swedish.

1. Introduction

Evaluation is a long-standing issue in natural-language processing, not least in machine translation. While the focus in recent years has been on metrics that can be computed automatically, such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006), they are not very informative. A potential user may be more interested in knowing the strengths and weaknesses of a given system. What can it do well? When is it likely to produce incorrect translations?

A common approach to more informative evaluations is error analysis (Vilar et al., 2006; Stymne and Ahrenberg, 2012). With a not too big and interpretable error taxonomy a user can get a good picture of what kind of mistakes a system is making when applied to a given text. A drawback with error analysis, though, is that the properties of the analysed text(s) are generally unknown so that we don't get information on what the system did right or on the frequency of constructions that sometimes give rise to errors.

Test suites may be seen as an alternative or complement to error analysis. King and Falkedal (King and Falkedal, 1990) discuss pros and cons of test suites for machine translation evaluation, and suggest that they can be valuable in spite of some pitfalls. One such list may be targeted at source language coverage, while another may be targeted at specific translation problems for the language pair in question, in particular at constructions where the two languages show 'mismatches'. They also argue that selection of inputs should be corpus-based.

When the paper by King and Falkedal was published, the capacity of a machine translation system was far below the capacities of present online systems. New technologies, such as Neural MT, are generally quite capable but also opaque and sometimes give errors that are hard to explain and describe. For this reason, (Isabelle, Cherry, and Foster, 2017) advocate a "challenge set approach" to evaluation of modern systems as a way to probe their capabilities. A challenge set, as the name suggests, should focus on difficult cases but the cases should be categorized so that the system output can be described and quantified in understandable terms. This paper presents a first version of a challenge set for English-Swedish machine translation.

A more ambitious approach to the use of test suites for machine translation evaluation is taken in the QT21 project (Burchardt et al., 2017; Macketanz et al., 2018). The ultimate goal is described as to "represent all phenomena relevant for translation" (Burchardt et al., 2017, p. 164) and provide for (semi-)automatic evaluation (Macketanz et al., 2018). Currently, their German-English test suite contains some 5,000 segments categorised into 15 major categories and some 120 different phenomena. The test suite is not published with the explicit reason "[t]o prevent overfitting or cheating"! .

2. The challenge set approach

In this work I decided to follow the challenge set approach centered around the notion of divergence or mismatch. A divergence is present in a translation if some construction in the source has been translated by a non-isomorphic construction. (Isabelle, Cherry, and Foster, 2017) suggests that a challenge set should be based on forced divergences, i.e., cases where the target language does not have an isomorphic construction, either because it does not exist at all in the language, or because the linguistic context is such that it cannot be used.

In (Isabelle, Cherry, and Foster, 2017) the divergences are divided into three major classes: morpho-syntactic, lexico-syntactic and (pure) syntactic ones. In (Isabelle and Kuhn, 2018) a fourth category, purely lexical divergences, was added. These categories are quite general, and not very informative in themselves. However, for the purposes of this paper, they are retained, and are defined as follows:

- **Morpho-syntactic divergence**. The divergence involves a morphological feature that is either not present in the source language, or, if it is, must change value in the target language sentence. A case in point for English-Swedish translation is gender agreement on determiners, pronouns, and adjectives. See Table 1 for an example.

- **Lexico-syntactic divergence**. The divergence involves a change in syntactic structure, such as complement structures, when a lexical item is translated by its typical synonym in the target language. For example, the English verb *want* is often constructed with an object NP and an infinitive VP (*want x to protest*) which is not available for the Swedish synonym *vill*, which instead requires a subordinate finite clause beginning with the subjunction *att*: *vill att x protesterar*.

SRC	The table she bought was **cheap**.
SYS	Bordet hon köpte var billig.
QUE	*Does the Swedish word translating 'cheap' have the proper form?*
SUG	billigt
ANS	YES NO NA

Table 1: A sentence with its focus question and suggestion.

- **Purely syntactic divergence**. The divergence involves a construction with no isomorphic counterpart in the target language. An example is the necessity to place a finite verb in the second position of a Swedish translation, although the English source verb may be in third or fourth position. Thus, an sentence such as *Kim seldom goes to the opera* cannot be translated with the same word order **Kim sällan går på operan*.

- **Purely lexical divergence**. These concern differences in selection of lexical items, including support verbs, prepositions and idioms. A case in point is the English verb *put* which usually requires a Swedish translation with a more specific sense.

An important aspect of the challenge set approach is that only one phenomenon for every example is evaluated. Every input sentence is supplied with a question, that focuses attention to some part of the source sentence, and that can be answered by a clear 'yes' or 'no'. For an illustration, see again Table 1.

Evaluation of a system with a challenge set is straightforward. The translations returned from a system are put into a form and each one is put together with its source sentence and the focus question. The human evaluators will then answer the question by yes or no. The performance of the system is captured by computing the share of correct translations in each category.

A challenge set can be used to compare several systems at one occasion or to compare different versions of the same system.

3. The English-Swedish Challenge Set

3.1 Design changes

We have made some minor changes to the design used by (Isabelle, Cherry, and Foster, 2017). They give a reference translation for each source sentence. As the question is focusing on a single aspect of the source, we believe that a complete reference translation may be too normative. Instead, the evaluator is given one or more suggestions for good translations of the focused part (SUG in Table 1), and, if known, responses that should be considered errors.

As in the English-French set every example carries a finer description of what divergence it is supposed to illustrate. In addition, the examples have been structured in pairs, with one member being judged a little more difficult to translate than the other. This added difficulty may have various sources, for instance, a longer distance between a targeted phrase and its governor, or the use of rarer words.

While the aim is to obtain a clear yes- or no-answer, this may not always be possible. For this reason the English-Swedish set includes a third alternative, NA, for 'not applicable'. This alternative can be used if the system somehow manages to circumvent the problem associated with the focused part, or if the evaluator cannot decide.

Category	Examples
Morpho-syntactic	48
Agreement in NP	10
ADJ-agreement in predication	14
Noun compounding	8
Pronoun coreference	6
Other	10
Lexico-syntactic	62
Sense-distinguishing context	30
NP-to-VP complements	8
Wh-phrases	8
Explicitation	8
Double object	4
Clauses with *fail to*	4
Purely syntactic	62
Word order	24
do-support	20
Inalienable possession	8
Clausal conjuncts	6
Tag questions	4
Purely Lexical	30
Sense specification	10
Idioms	20

Table 2: An overview of the data.

3.2 Contents

The current English-Swedish set contains 202 example sentences. The distribution on categories and phenomena is shown in Table 2.

Some sentences have been taken from the English-French challenge set, as they give rise to the same translation difficulty when translating into Swedish as they do for translating into French. Other sentences are made up but often based on sentences found in corpora that can be searched online such as the COCA corpus (Davies, 2018), the BYU-BNC (Davies, 2018), and the English-Swedish parallel UD corpora such as LinES and PUD (Nivre et al., 2018).

3.3 Evaluation

A thorough evaluation has not been undertaken, but we have made two pilot studies to get indicative answers to the following questions: (1) Are the sentences really challenging for current systems, and (2) Will different evaluators

Systems	Yes	No	NA	Accuracy
Sys 1	10	9	1	0.50
Sys 2	9	10	1	0.45
Sys 3	9	11	-	0.45
Sys 4	10	9	1	0.50

Table 3: A pilot system evaluation using twenty randomly generated challenge sentences. 'Yes' means the translation is judged correct in the focused aspect, 'No' that it is not. Judgements by author.

Items	User1	User2	User3	User4	Author
Q 1	NO	YES	NO	NO	NO
Q 2	NO	YES	NO	NO	NO
Q 3	YES	YES	YES	YES	YES
Q 4	YES	YES	YES	YES	NA
Q 5	NO	YES	YES	YES	YES
Q 6	NO	NO	NO	NO	NO
Q 7	YES	YES	YES	YES	NO
Q 8	YES	YES	YES	YES	YES
Q 9	YES	YES	NO	YES	NO
Q10	YES	NO	NO	NO	NO
Q11	YES	NA	YES	NA	YES
Q12	YES	NO	NO	NO	NO
Q13	NO	NO	NO	NO	NO
Q14	NO	NO	NO	NO	NO
Q15	YES	YES	YES	YES	YES
Q16	YES	YES	NO	YES	YES
Q17	YES	YES	YES	NA	YES
Q18	YES	YES	YES	YES	YES
Q19	YES	YES	YES	YES	YES
Q20	YES	YES	YES	YES	YES

Table 4: A pilot user evaluation.

agree in their judgements? The first question was tested by randomly selecting 20 sentences from the full set and have them translated by four different online systems. The output has been judged by the author and is shown in Table 3.

To test the second question the output from one of the systems was given to four other people with Swedish as their mother tongue. They were given a web form and very little instruction to perform the task. The results are shown in Table 4. There was full agreement only on less than half (9/20) of the items which means that the agreements are not as strong as could be hoped. Using majority voting 17 judgements can be seen to agree with those of the author. Table 5 shows some of the examples causing disagreement.

This small evaluation indicates that evaluators require detailed instruction, and that it may be useful to provide instances of translations that should be judged as incorrect. Lists of such answers may also pave the way for automatic scoring. It also indicates, however, that the label 'challenge set' is appropriate.

The current challenge set is available at https://github.com/LarsAhrenberg/En-Sv_MT_ChallengeSet/. Contributions, in the form of comments, reviews and additions are very welcome.

References

A. Burchardt, V. Macketanz, J. Dehdari, G. Heigold, J-T. Peter and P. Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

M. Davies. 2018a. BYU-BNC. A Corpus based on the British National Corpus from Oxford University Press. https://corpus.byu.edu/bnc.

M. Davies. 2018b. The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. https://corpus.byu.edu/coca.

P. Isabelle, C. Cherry, and G. Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, (EMNLP '17), Copenhagen, Denmark, pages 2486–2496.

P. Isabelle and R. Kuhn. A Challenge Set for French → English Machine Translation. ArXiv:1806.02725v1 cs.CL.

M. King and K. Falkedal. 1990. Using Test Suites in Evaluation of Machine Translation Systems. In Proceedings of the 13th Conference on Computational Linguistics (COLING-90) - Volume 2, Helsinki, Finland, pages 211–216. doi = 10.3115/997939.997976.

V. Macketanz, A. Renlong, A. Burchardt and H. Uszkoreit. 2018. TQ-AutoTest An Automated Test Suite for (Machine) Translation Quality. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18), May, 7-12, Miyasaki, Japan. European Language Resources Association (ELRA).

J. Nivre et al. 2018. Universal Dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2837.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 2002, pages 311–318. doi = 10.3115/1073083.1073135.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of Association for Machine Translation in the Americas (AMTA'06).

S. Stymne and L. Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12, May 23-25, Istanbul, Turkey.

D. Vilar, J. Xu, L.F. D'Haro and H. Ney. 2006. Error Analysis of Machine Translation Output. In Proceedings of LREC06, Genoa, Italy, pages 697–702.

SRC	*[A] terrifying black lion crossed the road.*		
SYS	*en skrämmande svart lejon korsade vägen.*		
QUE	Does the translation of the marked determiner agree with its head noun?		
ANS	1 YES	4 NO	- NA
SRC	*Only two of the players seemed to be [ready].*		
SYS	*bara två av spelarna verkade vara redo.*		
QUE	Does the translation of the marked word agree with its head noun?		
ANS	4 YES	- NO	1 NA
SRC	*As always, the bastard [failed to] respond.*		
SYS	*som alltid misslyckades den jäveln att svara.*		
QUE	Is the meaning of the marked verb correctly rendered in the Swedish translation?		
ANS	4 YES	1 NO	- NA
SRC	*They told you to [put] the bottle on the table, didn't they?*		
SYS	*de sa att du skulle lägga flaskan på bordet, eller hur?*		
QUE	Is the meaning of the marked verb correctly rendered in the Swedish translation?		
ANS	3 YES	2 NO	- NA

Table 5: Responses to items Q1, Q4, Q7, and Q9. The YES answer to Q1 can be interpreted as an error or sloppiness. The author's choice of NA for Q4 is due to the fact that the word *redo* does not inflect, so the system's ability to handle agreement is not tested by this choice. However, the translation is correct. The difference in judgements to Q7 perhaps indicates that the author is more sensitive to anglicisms than the average academic; he would prefer a simpler, idiomatic translation such as *den jäveln svarade inte.* The final example, Q9, is syntactically correct and semantically perfectly possible, but the most common posture of a bottle on a table is upright rather than horizontal and for that reason the answer NO is (perhaps more) adequate.