Semi-Automatic Evaluation of the Grammatical Coverage of Machine Translation Systems

A. Guessoum **&** **R. Zantout**

Dep. of Computer Science
The University of Sharjah
P.O. Box 27272
Sharjah, UAE
e-mails: guessoum@sharjah.ac.ae

Dep. of Computer Science
Faculty of Sciences
The University of Balamand
P.O. Box 100, Tripoli, Lebanon
rached.zantout@balamand.edu.lb

Abstract

In this paper we present a methodology for automating the evaluation of the grammatical coverage of machine translation (MT) systems. The methodology is based on the importance of unfolded grammatical structures, which represent the most basic syntactic pattern for a sentence in a given language. A database of unfolded grammatical structures is built to evaluate the parser of any NLP or MT system. The evaluation results in an overall measure called the grammatical coverage. The results of implementing the above approach on three English-to-Arabic commercial MT systems are presented.

Keywords

Machine Translation Evaluation; Translation Quality; Grammatical Coverage; Arabic Machine Translation; Parser Evaluation

I. Introduction

With the ever-growing interest in natural language processing (NLP) systems and machine translation (MT) systems, grows a need for developing adequate methodologies and tools for assessing their qualities. Such systems need to be assessed by system developers for any possible technological improvements and novel research ideas and by potential users for quality comparison purposes.

The evaluation of NLP systems is classically divided into two main approaches: *Glass-box* and *Black-box* (Hutchins & Somers, 1992), (Nyberg & Mitamura & Carbonell, 1993), (Arnold, 1995). In black-box evaluation, the evaluator has access only to the input and output of the system under evaluation. In Glass-box evaluation, the evaluator also has access to the various workings of the system and can thus assess each sub-part of the system. Component-based evaluation and detailed error analyses are also important types of evaluation (Nyberg & Mitamura & Carbonell, 1993), (Arnold, 1995), and (Hedberg, 1995). In addition, a NLP system, just like any other software, needs to be evaluated for its cost, performance, stability, maintainability, and portability among other criteria.

Subjectivity of NLP/MT system evaluation stems from the fact that evaluation methods rely heavily on humans. This makes the assessment depend on the evaluator's background, skills, or even taste! The evaluation process is affected by the degree of proficiency of the human evaluator in the various facets of the languages that are involved in the translation. Even when strict and clear rules are introduced at the beginning of the evaluation process, the exact score given to a system component under evaluation would vary from one (human) evaluator to another. Hence more objective evaluation methodologies need to be developed.

In Section II, a brief overview of some of the work on MT evaluation is given. In Section III, we introduce our methodology for evaluating the grammatical coverage of MT systems, which is similar to the methodology introduced in (Guessoum & Zantout, 2001) for MT system lexicon evaluation. This methodology minimizes the amount of subjectivity in the evaluation process by automating some evaluation tasks. Section IV presents and discusses the results of the implementation of the methodology for evaluating various AMT systems. Related work is presented in Section V, followed, in Section VI, by a summary of our contribution.

II. MT System Evaluation Efforts

(Hutchins & Somers, 1992) present a good survey of the various kinds of evaluations, namely: quality assessment in terms of accuracy, intelligibility, style, error analysis, and benchmark tests. (Lehrberger *et al.*, 1998) and (Dyson *et al.*, 1987) talk about the evaluation by users and (Melby, 1988), (Nagao, 1985) and (King *et al.*, 1990) on methodologies for MT evaluation. A number of methods for MT system evaluation are presented in (Vasconcellos, 1988). (Mellish *et al.*, 1998) explain how the problems of natural language generation are different from the problems of evaluating work in natural language understanding. An entire issue of the Machine Translation Journal was devoted to the evaluation of MT systems (Arnold & Humphreys & Sadler, 1993).

Notable evaluations of MT systems are those of Systran (Van Slype, 1979), and of Logos (Sinaiko & Klare, 1973). Major projects exist for the development of diagnostic and evaluation tools for Natural Language Processing applications, such as the DARPA project (White, 1994), the project DIET (Klein *et al.*, 1998) at DFKI (Germany), and the European project Eagles (Bevan *et al.*, 1998). Results of evaluating a large set of MT systems are found

in (Mason & Rinsche, 1995). There seems to be an agreement that the aspects of a MT system that should be evaluated are: *adequacy* (e.g. well-formedness and grammatical correctness) (White, 1994), *informativeness*, and *intelligibility* (Arnold, 1995).

The evaluation of Arabic MT/NLP tools is still very shy and very much non-systematic. (Jihad, 1996), (Qendelft, 1997), (Arabuter, 1996) present very brief surveys of a number of Arabic MT systems including Transphere, Arabtrans, and Al-Wafi. Such evaluations did not rely on any firm or formal evaluation methodology; however, they have shown some serious shortcomings of the MT systems.

III. Evaluation Methodology

In our work, we have chosen the black-box evaluation approach due to the fact that we want to evaluate commercial systems and, consequently, we have no access to their inner workings. Even so, it is desirable to be able to draw from such an evaluation enough conclusions about the various system components. In such a setting, the evaluation may not be able to pinpoint the error source, however it will give an indication as to what sub-system is malfunctioning.

Similar to word senses in the methodology presented in (Guessoum & Zantout, 2001), unfolded grammatical structures are used here. Table 1 shows sentences with their unfolded grammatical structures. These are the most basic grammatical patterns which, once instantiated using entries from the lexicon, produce sentences of the language.

Sentence	**Unfolded Grammatical Structure**
He is walking.	pronoun (3 sing., masc) + aux. (3, sing.) + verb (present, prog.) : active voice
He was walking.	pronoun (3, sing., masc) + aux. (3, sing.) + verb (past, prog.) : active voice
John laughed.	proper_noun + verb (past, simple) : active voice
John found a key.	proper_noun + verb (past, simple)+ det. (sing.) + noun(sing.) : active voice
You can see the house.	pronoun (Number1) + modal + verb (inf-to) + det.(Number2) + noun(sing.) : active voice
I will have seen the house.	pronoun (1 sing.) + modal (future)+ aux. + verb (past_participle) + det.(Number) + noun(sing.) : active voice
My hat will be hidden in the drawer.	possessive_pronoun (1 sing.) + noun + modal(future) + aux. + verb (past_participle) + prep. + det.(Number) + noun(sing.) : passive voice
Whose books did you find?	wh-determiner + noun (plural) + aux. (past) + pronoun (Number) + verb (inf-to): active voice
The men we saw at the store are intelligent.	det.(Number) + noun (plural) + pronoun(1 plural) + verb (past, simple) + prep. + det. (Number) + noun(sing.) + aux (present, plural) + adjective : active voice

Table 1: Examples of different sentences and their unfolded grammatical structures

The following two sentences are grammatically equivalent, since they can be obtained by instantiating the same unfolded grammatical structure:
The men we saw at the exhibition are intelligent.
The toys they bought in the supermarket are interesting.
However, the same two sentences are not grammatically equivalent to any of the following sentences:
The men whom we saw at the exhibition are intelligent.
The lions that ate the horse were hungry.
The toys we took from the supermarket to the car are expensive.

Note that sentences (1) and (3) above are semantically, but not grammatically, equivalent. Indeed, (3) can be generated by instantiating the same unfolded grammatical structure that allows to generate (1).

Definition: A parser/generator *covers* an unfolded grammatical structure if the latter can correctly be analyzed by the former.

Definition: The *grammatical coverage* of a parser/generator is defined as the ratio of the number of unfolded grammatical structures that this parser covers to the total number of known unfolded grammatical structures.

Obviously, the automation of the grammatical coverage evaluation process requires the existence of a tool for parsing input sentences to find out the unfolded grammatical structure of each input sentence. It also requires a tool to check the MT system ability to translate each of these unfolded grammatical structures correctly.

For a correct assessment of the parser's grammatical coverage, the evaluation should be done independently of erroneous behavior of any other system component. For instance, the grammatical coverage should be computed using sentences that do not contain any errors due to improper system lexical coverage.

Grammatical coverage evaluation can be done using one of two methods.

Method I

This method relies on treating all the unfolded grammatical structures equally. This approach is more suitable when no statistics are available as to the relative importance of unfolded grammatical structures. The grammatical coverage of the MT system can thus be computed using Equation (E1).

$$Grammatical\_Coverage(MT) = \frac{\sum_{i=1}^{N} d(G_i, MT)}{N} \quad \text{(E 1)}$$

where:

MT is the machine translation system which needs be evaluated.

Gi is the unfolded grammatical structure *i*.

d(Gi,MT) = 1 if *Gi* was handled completely correctly by the MT system

if *Gi* was handled partially correctly

0 if *Gi* was handled incorrectly.

N is the number of language unfolded grammatical structures (the number of test structures).

Method II

Here the evaluator assigns a different weight for each unfolded grammatical structure of the language. This weight reflects the statistical occurrence of such a structure in the set of test sentences (database). The idea is a straightforward extension of the notion of word sense weight (Guessoum & Zantout, 2001) for lexicon evaluation.

$$W(G_i) = \frac{Occurrences(G_i)}{\sum_{j=1}^{N} Occurrences(G_j)} \quad \text{(E 2)}$$

where:

Gi and *N* are defined as above,

Occurrences(Gi) is the number of occurrences of the unfolded grammatical structure *i* in the language (test sentences in our case), and

W(Gi) is the weight of the unfolded grammatical structure *Gi*.

The grammatical coverage of the MT system is then calculated using Equation (E3).

$$Grammatical\_Coverage(MT) = \sum_{i=1}^{N} W(G_i) * d(G_i, MT) \quad \text{(E 3)}$$

The main advantage of Method II is that it gives more weight to the grammatical structures which are commonly used in a given language. Therefore, the systems that handle such grammatical structures adequately will have a better rank than those that do not.

Procedures *build_grammar_structures_db* and *evaluate_parser* (given below) implement method II (method I is a special case).

Procedure *build_grammar_structures_db*
input:
- text in the source language (corpus)

output:
- a database of unfolded grammatical structures and their occurrence rates.

Begin
While not end of text
Do
- Read a sentence from the text
- Analyze it recognizing its unfolded grammatical structure
- Increment the number of occurrences for the corresponding unfolded grammatical structure in the DB

EndWhile

- update the DB statistics using formula E2.

End.

Procedure *evaluate_parser*
input:
- a DB of unfolded grammatical structures and
- statistics (built using procedure *build_grammar_structures_db)*
- a MT system

output:
- grammatical coverage of the parser

Begin
Sum = 0

- **For** each unfolded grammatical structure g_i in DB
Do
If g_i can be handled by the MT system parser
Then $d(g_i, MT) = 1$
Else $d(g_i, MT) = 0$

Sum = Sum + $d(g_i, MT)$
EndFor

Grammatical_coverage(MT) = Sum
(equivalent of formula E3)

End.

IV. Evaluating Arabic MT Systems

Our MT system evaluation methodology was tested on the following commercial English-to-Arabic MT systems that we have managed to purchase:

Al-Mutarjim Al-Arabey was produced by ATA Software Technology Limited. The company claims that it

is "the first English-Arabic MT System ever to be developed on personal computers" (ATA, 1997). system contains comprehensive dictionaries (300,000 "lines of words"), a good level of "text context analysis", the introduction of different word senses, whenever available, and a correct translation of most of the common abbreviations.

Al-Wafi was also developed by ATA Software Technology Limited. From our evaluation, we concluded that it uses the same MT modules as Al-Mutarjim Al-Arabey, except for a less extensive lexicon.

Arabtrans was developed by Arab.Net Technology limited. According to its developers, Arabtrans translates texts from English to Arabic at more than a thousand words per minute but "…the translation produced by the program requires editing for both grammatical accuracy and to check whether alternative meanings are preferable" (Arab Net, 1996).

We have not been able to purchase a fourth system, Transphere (by Apptek), despite multiple attempts and direct contacts with the company representatives.

According to (ATA, 1997) and (Al-Jundi, 1997), the

Sample texts for assessing a parser grammatical coverage were chosen so that at least one test case exists for the most important grammatical structures. The scoring, by a human, of the output Arabic sentences was done as follows:

- A score of 1 if the sentence grammatical structure is correct with a clear meaning.
- A score of 0.5 if there is something missing or incorrect such as diacritization, case endings, and pronouns, but the grammatical structure is roughly correct.
- A score of 0 if the sentence grammatical structure is completely incorrect.

Table 2 shows the results obtained, assuming all unfolded grammatical structures have the same weight (Method I). These results show a weakness in the grammatical coverage for all three systems. The maximum grammatical coverage percentage is 57.5%, a low score that explains the frequently bad output quality.

	Al-Mutarjim Al-Arabey	**Arabtrans**	**Al-Wafi**
Coverage of Grammatical Structures	57.5 %	32 %	57.5 %

Table 2: Results of Grammatical Coverage Evaluation of AMT Systems

Table 3 details the results more clearly with respect to various classes of unfolded grammatical structures. All three AMT systems perform poorly on each of the grammatical categories. Al-Mutarjim Al-Arabey and Al-Wafi have performed exactly the same for all categories, which indicates that they use the same parsing engine. On the other hand, Arabtrans has scored a very low overall grammatical coverage of 32 % and is better only in simple Tense structures.

Grammatical Structure Form	**Al-Mutarjim Al-Arabey**	**Arabtrans**	**Al-Wafi**
Verb Forms	75 %	35 %	75 %
simple tenses	58 %	83 %	58 %
Progressive tenses	58 %	17 %	58 %
Various forms of conjunction	50 %	40 %	50 %
Noun phrase and Verb phrase combinations	69 %	31 %	69 %
Different combinations	89 %	56 %	89 %
Auxiliary verbs	50 %	16 %	50 %
Active voice sentences	83 %	17 %	83 %
Passive voice sentences	33 %	25 %	33 %
WH-Questions	39 %	18 %	39 %
Relative clauses	17 %	8 %	17 %

Table 3: Detailed grammatical coverage evaluation for 3 AMT systems

V. Related Work

One finds in (Van Slype, 1979) a summary of a number of approaches that calculate various statistics about various MT system features. Miller and Beebe in (Halliday & Briss, 1977) establish an *a-priori* list of syntactic constructions; take the results of human translation (HT) and MT; and calculate the ratio of the number of constructs common to the MT and HT versions over the total number of occurrences of the syntactic constructions in the HT version. Our approach is quite different in that it takes into account the weights of each syntactic construct and whether it is translated into the appropriate grammatical construct of the target language by the MTsystem. Weissenborin (Van Slype, 1979) suggests a syntactic evaluation based on the ratioof the number of the source language analysis grammar rules existing in the MT system to the number of grammatical rules in the source language for the type of texts to be treated. Thus, the grammatical coverage in Weissenborn's work is defined only in terms of the source language, whereas we take into account the coverage by the MT system of the source language modulo the weights of its unfolded grammatical constructs as well as the appropriateness of the target grammar construct produced by the MT system. The same comments can be made about the differences between our approach to grammatical coverage evaluation and that adopted in (Chaumier & Mallen & Van Slype, 1977) where a finer scrutiny is done of the grammatical (sub-) constructs in the texts.

In (Van Slype (report), 1979) a detailed study of the methods that had been developed for evaluating machine translation is presented. The report subdivides the evaluation features into two main categories: macro-evaluation and micro-evaluation. Relevant to our paper are the micro-evaluation methods which can be subdivided into five groups that include the grammatical symptomatic level, i.e. the analysis of the grammatical errors found in the target output.

In (Carroll (ed.), 1998) a number of papers on the evaluation of parsing systems are included along with a survey of parser evaluation methods (Carroll & Briscoe, 1998). The methods were divided into corpus-based and non-corpus based. Listing the linguistic constructions covered by a particular parser is a non-corpus based approach. Relevant un-annotated corpus-based methods calculate (1) the percentage of sentences assigned one or more analyses by a parser; (2) the geometric mean of the number of analyses divided by the number of input tokens in each sentence parsed; and (3) a measure of the degree to which a probabilistic language model minimizes unpredictability and ambiguity. . Relevant annotated corpus-based methods calculate (1) the percentage of sentences which receive one or more analyses that are consistent with the correct analysis in the corpus. (2) the percentage of highest-ranked analyses output by a probabilistic parser which are identical to a manual analysis provided in an annotated test corpus (tree bank). Other approaches use tree similarity measures of various types, etc.

VI. Conclusion

We have introduced in this paper a methodology for evaluating the grammatical coverage of MT system components in a black-box setting. The methodology is a generalization of that for MT system lexicon evaluation of (Guessoum & Zantout, 2001) and is based on the concept of unfolded grammatical structure and its occurrence frequency. The result is an assessment of the ability of a parser to handle various grammatical structures. The methodology presented in this paper, while being useful for systematic evaluation of Arabic MT systems, is general enough to be applicable to the evaluation of any MT system.

The evaluation methodology was implemented and tested with three Arabic MT systems. The evaluation results have shown a poor grammatical coverage for all three systems, confirming the disappointing output frequently produced by such systems. The evaluation has also shown that two systems, produced by the same company, use the same parsing engine.

The grammatical coverage evaluation can be improved by implementing "Method II" of Section III.3. As a prerequisite, a tool which parses an input sentence and returns the corresponding unfolded grammatical structure should be developed. With the help of this tool, statistics about the occurrence ratios of the various unfolded grammatical structures should be computed. The methodology can also be extended to the evaluation of the semantic correctness, pronoun resolution correctness, and style of translated text (the latter being sensibly more complex).

Acknowledgements

We would like to thank MSc student A. Al-Sikhan who implemented the methodology introduced in this paper.

Bibliographical References

Al-Jundi, F. (1997). Al-Mutarjim Al-Arabey: An Attempt to Understand English. PC Magazine (Middle East), October,. (pp. 40—44). (In Arabic)

Arab. Net. Technology. Ltd.. (1996). Arabtrans User's Guide. Arab Press House.

Arabuter. (1996) al-mutarjim al-aaliy al-waafy [The Machine Translator: Al-Wafi]. Arabuter. 8 (71), September.

Arnold, D.J. (1995). Evaluating MT Systems. December. http://c1www.essex.ac.uk/~doug/book/node75.html.

Arnold, D.J., & Humphreys R.L. & Sadler L. (Eds.). (1993) Machine Translation Journal: Special Issue on Evaluation of MT Systems. (1-2), Kluwer Academic Publishers.

ATA Software Technology. (1997). Al-Mutarjim Al-Arabey User manual. http://www.atasoft.com

Bevan, N. et al. (1998). Proceedings of the second EAGLES II Workshop on Evaluation in Human Language Technology. Geneva, 8-9th September.

Carroll, J. (Ed.). (1998). Proceedings of the Workshop on the Evaluation of Parsing Systems. University of Sussex, CSRP 489, June.

Carroll, J., & Briscoe, T. (1998) A survey of Parser Evaluation Methods, in the Proceedings of the Workshop on the Evaluation of Parsing Systems. Carroll (ed.), J. University of Sussex, CSRP 489, June.

Chaumier, J., & Mallen, M.C. & Van Slype, G. (1977) Evaluation du Système de Traduction Automatique SYSTRAN: Evaluation de la Qualité de la Traduction. CEC Report N0 4, June. Luxembourg.

Dyson, M.C. & Hannah, J. (1987). Towards a Methodology for the Evaluation of Machine-Assisted Translation Systems, Computers and Translation. 2, (pp. 163—176).

Guessoum, A. & Zantout, R. (2001). A Methodology for a Semi-Automatic Evaluation of the Lexicons of Machine Translation Systems. to appear in the Machine Translation journal, Kluwer Academic.

Halliday, T.C., & Briss, E.A. (1977). The Evaluation and Systems Analysis of the Systran Machine Translation System. NTIS ADA 036.070, January.

Hedberg, S. (1995). Machine Translation Comes of Age. Computer Select. September.

Hutchins, John & Somers, H.L. (1992). An Introduction to Machine Translation. Academic Press.

Jihad, A. (1996). hal bada'a `aSru altarjamati al-aaliyyati `arabiyyan? [Has the Arabic Machine Translation Era Started?]. Byte Middle East. November. (in Arabic).

King, M. & Falkedal, K. (1990). Using Test Suites in Evaluation of Machine Translation Systems. Proceedings of the International Conference on Computational Linguistics. (pp. 211-216).

Klein, J., & Lehmann, S. & Netter, K. & Wegst, T. (1998). DiET in the Context of MT Evaluation. KONVENS98. 5-7 October.

Lehrberger, J. & Bourbeau, L. (1998). Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation, Linguisticae Investigationes, 15.

Mason, J. & Rinsche, A. (1995). Ovum Evaluates: Translation Technology Products. OVUM ltd., June.

Melby, A.K. (1988). Lexical Transfer: Between a Source Rock and a Hard Target. Proceedings of the International Conference on Computational Linguistics, (pp. 411-419).

Mellish, C. & Dale, R. (1998). Evaluation in the Context of Natural Language Generation. Journal of Computer Speech and Language, 12, 349--373.

Nagao, M. (1985). Evaluation of the Quality of machine-Translated Sentences and the Control of Language. Journal of the Information Processing Society of Japan, 26 (10), 1197--1202.

Nyberg, E.H. 3rd. & Mitamura, T. & Carbonell, J.G. (1993). Evaluation Metrics for Knowledge-Based Machine Translation. Center for Machine Translation, Carnegie Mellon University, PA. http://www.lti.cs.cmu.edu/Research/Kant.

Qendelft, G. (1997). barnaamaj alwaafy liltarjamati mufiidun lifahmi alma`naa al`aammi min risaalatin inkliiziyyatin. [The Translation Program Al-Wafi Is Useful for Getting a General Understanding of a Letter Written in English]. Al-Hayat newspaper, (12657), 25 October, (in Arabic).

Sinaiko, H. W. & Klare, G. R. (1973). Further Experiments in Language Translation: A Second Evaluation of the Readability of Computer Translations. ITL, 19, 29--52.

Van Slype, G. (1979). Systran: Evaluation of the 1978 Version of the Systran English-French Automatic System of the Commission of the European Communities. The Incorporated Linguist, 18, 86--89.

Van Slype, G. (1979). Critical Study of Methods for evaluating the Quality of Machine Translation, (Final Report). Prepared for the Commission of the European Communities. Brussels.

Vasconcellos, M. (Ed.). (1988). Technology as Translation Strategy. American Translators Association Scholarly Series, 2.

White, J. & O'Connell, T. & O'Mara, F. (1994). Machine Translation Program: 3Q94 Evaluation. Advanced Research Projects Agency, (http://ursula.georgetown.edu/mt_web/3Q94FR.htm).