Evaluation of Linguistics-Based Translation

Janne Bondi Johannessen*, Torbjørn Nordgård, Lars Nygaard***
University of Oslo*, LingIT**
The Text Laboratory, ILN, P.O.Box 1102 Blindern, N-0317 Oslo
jannebj@iln.uio.no, torbjorn@lingit.no, larsnyg@iln.uio.no

Abstract

We report on the evaluation of the Norwegian–English MT prototype system LOGON. The system is rule-based and makes use of well-established frameworks for analysis and generation (LFG and HPSG). Minimal Recursion Semantics is the "glue" which performs transfer from source to target language and serves as the information vehicle between LFG and HPSG. The project-internal testing uses material from the training data sources. We report on two test methods in addition: 1) on test data from the same (narrow) domain as the project was designed for, and 2) on a syntactic test suite. The former turns out to give much worse results than the project-internal tests, while the latter obtains somewhat better results. These results are important for future evaluations of similar systems.

1. Introduction

1.1 The LOGON MT system

In this paper we report on the evaluation of the MT prototype system LOGON, developed for Norwegian–English translation. LOGON is a joint project involving three Norwegian universities, aiming at a "deep" approach: to deliver high-quality MT based on the combination of a symbolic, semantic-transfer-oriented backbone and stochastic processes for ambiguity management and robustness (Oepen et al. 2004).[1] At the heart of the system is Minimal Recursion Semantics (MRS; Copestake, Flickinger, Sag, & Pollard, 2006) together with formalized grammatical frameworks for analysis and generation (LFG and HPSG). The textual domain is "guidebooks for mountain hiking in the summer season in Southern Norway".

1.2 Evaluation

The paper describes and discusses evaluation of the specific "Fjell" release (of January 2007) of the LOGON system. The project decided that "official" testing should use the training material and held-out texts from the same sources. The original work plan anticipated a 95% success rate on the training material (500 sentences) and an 80% success rate on unknown texts (from known sources) at the end of the project period (LOGON Work plan for 2005 and 2006). A less ambitious goal was defined after the first two years, to provide translations for 50 % of the sentences in the test material taken from the same domain. There was no explicit accuracy goal, so we have had to interpret the goal w.r.t. our own evaluation scales (see section 2).

In addition to the project-internal evaluation we also report on results from evaluations of the system on two other kinds of material: 1) novel test data from the same domain, and 2) a syntactic test suite. The results of 1) show that a linguistically based system achieves much worse results on test data from a different text source than the training material, even in the same narrow domain. The results of 2) show that the syntactic test (controlled for vocabulary), yields better results. Thus, future evaluations of similar systems should include both kinds of test material.

The motivation for test 1) is supported in the literature: Tessiore and Hahn (2000:615) report on Verbmobil, also a deep linguistics-based MT system: *"In general, in any validation or evaluation procedure [...] the independence of the test from the development is a crucial feature. Only in very special algorithmic cases can a finite and closed number of system states be used for both, development and test. Linguistic processing, in contrast, is a completely different case, because every linguistic theory can even prove that the linguistic performance creates infinitely many surface structures from the limited structural material."*

We believe this is a very insightful remark. A hand-crafted linguistics-based system will be able to generalise over input given in the training process. For example, if a verb occurs in the infinitive in the training corpus, the developer will immediately add the other forms of that verb to the lexicon of the system. Generally, a word used in one meaning in the training corpus may inspire the developers to add similar words and their arguments and syntax into the system. Statistically trained systems are blind to these kinds of generalisations. Therefore, to test the generality of a linguistics-based, hand-crafted system, it is important to use a new text, written by other authors than those of the training material.

[1] The present authors – the evaluation team – have not been part of the central developing team, although we have been part of the LOGON project team and have performed occasional tasks for the developers.

The motivation for a syntactic test (test 2) is that the LOGON system is linguistics-based. Much effort has been put into the grammatical qualities of the system. For this reason, it would be unfair not to test this quality separately. A systematic syntax test reveals what kinds of constructions the system can deal with and which it cannot. Since the system is sensitive to vocabulary issues, the syntactic test suite contains only words that have been tested repeatedly so that we know that whenever a syntactic construction is not given a translation, it is not the fault of any particular vocabulary item missing.

While automatic evaluation methods have become popular for statistical MT, e.g. BLEU (Papineni et al. 2001), we find it important that evaluation of a linguistics-based system must focus on fidelity and fluency (Hovy, King and Popescu-Belis 2002). For the project-internal testing we had an experimental set-up with eight external assessors. The assessors had native English competence, and had higher education, but were not linguists. We used three different LOGON "systems" differing thus: i) TOP: Translations selected as best by the LOGON system itself. The TOP system chooses the best candidate by means of statistical ranking for each module (analysis, transfer, generation), in a combination of n-grams and maxent including linguistic features. See Velldal et al. (2005), Velldal (2007) and Oepen et al. (2007). ii) ORACLE: Translations that are judged as best by a human oracle. The ORACLE was an American speaker with fluent knowledge of Norwegian, and was one of the development team. She was instructed to pick, from each set of translations, the one she liked best w.r.t. fidelity and fluency. iii) BLEU: Translations deemed best according to the BLEU metric (three reference translations were available for all test sentences).

2. The evaluations

2.1 The "official" testing

The internal test material (from three hiking guides) was divided between training material and held-out data from the same sources. The developers had access to a full vocabulary list of half of the held-out material. The amount of unknown words in the other half need not be high because the test sentences are collected from the same sources as the training material.

Out of 446 sentences in the test set, 254 sentences were translated by the system, i.e. 57 %. The numbers are broken down w.r.t. known-ness type in table 1.

	Training set (e)	Unseen sentences, known vocab (k)	Unseen sentences, unknown vocabulary (u)	Total
No. tested	144	144	158	446
No. translat.	86	89	79	254
% translated	60 %	62 %	50 %	57 %

Table 1. Number and per cent of translated sentences in the three source books.

The 57 % translated sentences were put into a web-based evaluation system, in which the eight external assessors would judge the translation by grading fluency and fidelity, and writing comments. For a given translation, the full report of the eight translators looks like in figure 1. The grades go from 0 to 3, where 2 and 3 are good grades. We are aware that some researchers like to separate fluency and fidelity, so that they are not evaluated at the same time by the human assessors (see Hamon et al. 2007). However, we chose to present them together.

Each assessor was given five translations of each sentence. Ideally, there would be one each of the TOP, ORACLE and BLEU systems, plus two simple n-gram trained, experimental reference systems, SMT and OA. OA (Oversettelsesassistenten ("The translation assistant"), developed at NTNU in Trondheim), was trained on general material, and SMT, a straightforward SMT implementation (GIZA ++, Pharaoh; Koehn 2004a, 2004b), was trained on the same training set as LOGON used plus a Norwegian-English parallel corpus available at the University of Oslo. However, since the three LOGON systems sometimes gave the same output, two human translations served as reserves to be put in. The translations were presented in random order, and the assessors had no idea which translation came from which source.

50062: Herfra går det 1500 meter rett ned til bunnen av Sunndalen.

a: **From here, 1500 meters goes directly down to the bottom of Sunndalen.***{tgu : BLEU}*

user	flu	fid	gram	voc	miss
2	0	3	wrong word orderno comma1500 meter = plural --> 'go', not 'goes'A meter cannot go anywhere	'it is ... straight down into the ... of Vally Sunndalen', not '1500 meters goes directly down to the bottom of Sunndalen'.	Valle (Sun
3	2	2	wrong word order, goes wrong verb form		it
5	1	1	wwo"it" after comma	"goes" => "drops"	
7	3	3			
8	2	2		wrong word choices	
9	3	2	goes=go	directly = I prefer straight.	
12	3	3			
13	2	2	From here, 1500 meters goes directly down to the bottom -> From here, it goes down 1500 meters straight to the bottom		

Figure 1: Screen shot of part of the window showing assessors comments, for experimenters. Abbreviations: flu=fluency, fid=fidelity, gram=grammar, voc=vocabulary, miss=missing items. (Idiomacity is also a box, but not visible in this figure.)

Below, we see the assessors' grading of the 57 % that were translated by the LOGON "systems", and the equivalent translations by the two statistical MT systems SMT and OA (which were only evaluated for the sake of comparison, and not in their own right).
The grading scale is defined as follows:

	FIDELITY	FLUENCY
0	Useless	No coherent English
1	Some	Something is OK
2	Fair	Still some mistakes
3	Perfect	Perfect English

	Fidelity				
	ORACLE	BLEU	TOP	SMT	OA
Train	2,27	2,11	2,03	2,12	1,16
Known	1,96	1,82	1,68	1,53	1,24
Unknown	2,16	2,08	2,01	1,65	1,33

Table 2: Fidelity measures.

	Fluency				
	ORACLE	BLEU	TOP	SMT	OA
Train	1,93	1,80	1,73	1,68	1,29
Known	1,73	1,61	1,52	1,26	1,26
Unknown	1,87	1,78	1,75	1,37	1,28

Table 3: Fluency measures.

As expected, the ORACLE selects the best translations. Something that at first glance looks very surprising is the fact that for both fidelity and fluency, the scores for test material with possibly unknown vocabulary are higher than those for known vocabulary. However, the length of the sentences with possibly unknown vocabulary is on average shorter than that of the known vocabulary ones (7:9 words for the main test text). Also, it should be remembered that the numbers only include sentences that actually have been translated, so that there is in fact no unknown vocabulary in the translated sentences. Still, there is almost the same number of sentences that have been translated in the known and unknown vocabulary test set (60:55 sentences for the main text). Thus, when the difference is not higher between the two sets, our initial motivation for including a new test set with text by new author must be said to be justified (section 1.2). The similarity between them shows that there cannot be very much that is really unknown in the possibly unknown test set.

The BLEU metric is clearly better than LOGON's selection mechanism TOP (significantly better for most categories, see Johannessen, Nordgård and Nygaard, to appear), showing that the use of reference translations is indeed a powerful tool compared to the system's own selection metric. Both tables show that fidelity and fluency are good for the 57 % sentences that LOGON actually translated (i.e. the ORACLE, BLEU and TOP "systems"). The internal testing is presented in full in Johannessen, Nordgård and Nygaard (to appear).

An advantage of using human assessors in the evaluation process, is the possibility of getting comments to the translations, as exemplified in table 4. In retrospect we think it might have been better if the comments had been given via a strict questionnaire (see e.g. Elliott et al. 2004). Then their comments would have been possible to study automatically. Instead, we chose to let the actual grades be automated, while the comments could be written freely. However, a questionnaire with all the answer alternatives given would have required a lot of experience that we did not have at the time. The comments give us valuable knowledge about what assessors notice and how they deal with it, and this is something we can use later.

Translation by TOP, ORACLE or BLEU	**Some comments by assessors**
Summer and winter Do marked routes go since Gjendesheim, and into Gausdal Vestfjells?	horrible syntax
	wrong form and placement of verbs
	This is absolutely gibberish!
Buses don't go through Kviknebygda into the weekends.	'during' or 'at', not 'into'
	"into" => "at"
	into the=during
	into -> on
	wrong preposition

Table 4: Some translations and the comments given by some assessors.

2.2 Evaluation material from different sources

As mentioned, it is important when testing a linguistic processing system that test data are taken from independent sources (Tessiore and Hahn 2000). We have thus tested the LOGON system on data from another source, in the same narrow domain: *Til fots i Jotunheimen (Hiking in Jotunheimen)*. Here, we have chosen a system of two grades: OK or Fail. In the first category, we accept any sentence that preserves the meaning even if it is somewhat flawed grammatically (grade 2-3). The Fail category refers to sentences that have not been translated or have a wrong meaning. In practice it can be said to be a fidelity measure, although there was hardly any sentence in this test that was translated well w.r.t. meaning, but had low grammatical score. The evaluation team performed this task, using the LOGON TOP system.

	The Hardanger Plateau	Rondane National Park	Total
No. tested sentences	17	15	32
No. translated sentences	4	5	9
% translated sent.	24 %	33 %	28 %
No. OK sentences	4	3	7
% OK sentences	24 %	20 %	22 %

Table 5: The LOGON system tested on texts by new author.

We exemplify the data with three sentences, two OK and one Fail, respectively:

(1) (OK)
Source: Fem turistforeninger arbeider med å holde rutenettet og turisthyttene i orden på Hardangervidda.
LOGON: 5 travel associations work with || Keep the route network and the tourist cabins in an order at Hardangervidda
Reference: Five tourist associations work to keep the trail network and tourist cabins in order on the Hardanger Plateau

(2) (OK)
Source: Hardangervidda har den største villreinstammen i Europa.
LOGON: Hardangervidda has the largest wild reindeer herd in Europe.
Reference: The Hardanger Plateau has the largest stock of wild reindeer in Europe

(3) (Fail)
Source: Helt siden begynnelsen av forrige århundre har DNT tilrettelagt for vandring i Rondane.
LOGON: Some main person since the beginning of a last century, DNT has arranged for walking in Rondane.
Reference: Ever since the beginning of the previous century, DNT has laid the groundwork for hiking in Rondane.

It is often the case that vocabulary is the main obstacle for arriving at a translation. Below are some items that were given in the error output of the system:

|"_frede_v_rel"|.	protect
|"_halvpart_n_rel"|.	half
|"_skifer_n_rel"|,	shale, slate
|"_lava_n_rel"|,	lava, volcanic rock
|"_bergart_n_rel"|.	rock type

Table 6: Items that were returned from the LOGON system as untranslated.

Recall that with data from the same sources as the training data, the system translated 57 % of the sentences. With independent test data, 28 % were translated, i.e. a difference of 30 percentage points. It is clear that the present test has a vocabulary that is somewhat different from the project-internal test data. Importantly, this shows how different authors use different words even in a narrow domain, and that the performance of a MT system (indeed any text-oriented language technology system) should be tested on textual sources different from those used in the training phase.

2.3 Evaluating the syntax

For a linguistics-based system, it is natural to test it purely grammatically, without vocabulary items destroying the output. We have compiled a syntactic test-suite on the basis of an introductory syntax book (Lie 2003). It contains sentences covering 154 constructions, with examples of topicalization (subjects, objects, finite or infinite verbs, clauses), extraposition, simple and complex clauses, formal subjects, clefting, long-distance dependencies, etc. The evaluation team carried out the test, collapsing fidelity and fluency into one accuracy measure, with two grades: OK (grade 2-3) and Fail (0-1).

	Syntactic test sentences	Translated sentences
Number	154	135
Per cent	100 %	88 %

Table 7: Number of translated sentences in the syntactic test suite.

We see that 88 % of the sentences have beeen given a translation (good or bad). Recall that the demonstrator returned some translation (whether good or bad) to 57 % of the test material from the known sources, and 28 % to that from sources by new author. In this light, 88 % is good. Out of these, 63 % were judged acceptable. The acceptable sentences were thus 55 % of the total number of sentences in the syntactic test suite, see table 8.

	Good translations	Bad translations	Ambiguous translations	Total translated sentences	Good translations of total sentences
Number	85	48	2	135	154
Per cent	63 %	36 %	1 %	100 %	55 %

Table 8: Sentences with acceptable translations in the syntactic test suite.

The syntactic test suite covers many constructions, most of which occur in normal written prose. There are seven sentences that could possibly be said to be less frequent in written texts, i.e., dislocation. There are seven examples of dislocation in the test suite, including both nominal and adverbial dislocations. The achievement of 55 % therefore may seem a bit low. Four translations are exemplified below, two OK and two Fail, respectively. The first one tests whether the system manages a subject realized as a finite clause, and the second one whether it handles double objects with a lexical direct object. Both are OK.

(4) (OK)
Source: At bjørnestammen var tallrik, var storslått.
LOGON & Correct: That the bear population was abundant was magnificent.

(5) (OK)
Source: Han viste henne ikke stien.
LOGON & Correct: He didn't show her the path.

The first Fail example illustrates, however that object shift, i.e., when a pronominal direct object is to the left of the adverb, is not translated correctly; the pronoun is represented by a peculiar reflexive.

(6) (Fail)
Source: Han viste henne den ikke
LOGON: He didn't show her itself.
Correct: He didn't show her it.

Sentence (7) tests whether the system can handle a topicalized object with a simple verb. The source sentence is syntactically ambiguous: subject-verb-object and object-verb-subject. Semantically the sentence can be disambiguated easily since it is people, not huts, who can build things. However, the LOGON system only returns one output, the wrong one.

(7) (Fail)
Source: Seterbuene reiste folk vinterstid.
LOGON: The mountain farm huts set up people in the winter.
Correct: People set up the mountain farm huts in the winter.

3 Conclusion

We have carried out three kinds of evaluations. The main, "official", evaluation was an experimental set-up based on text from the same sources as the training corpus. This test material consisted of a) training texts, b) test texts with known vocabulary and c) test texts with possibly unknown vocabulary. 446 sentences were tested, and 254 sentences were translated by the system, i.e. 57 %.

The experimental set-up consisted of eight human assessors who evaluated three translations of each of the 254 sentences using a web-based evaluation system. These sentences were given a scale from 0 to 3 for both fidelity and fluency. The best translation that the LOGON system chose from its many candidate translations, the TOP system, received an average of 1.83 for fidelity, and 1.62 for fluency. The system has potential for improvement, though: The translation candidate picked by a human oracle got an average of 2.05 and 1.79, respectively. Recall that the grade 2 for fidelity means that "Much of the meaning of the source sentence is transferred". The TOP system is approaching that, while the ORACLE system has managed it. The grade 2 for fluency means: "The sentence has many good features, but something is still wrong." Unfortunately, neither system has reached this goal.

Judging now whether the demonstrator has reached the project goal of translating more than half of the test material, the answer is positive, given the 57 % translated sentences. However, the quality numbers for fidelity and fluency for the translated sentences show that the quality can be discussed. The original goal of 80 % translations is far from having been reached.

We have also carried out two smaller evaluations; one with test material from the same domain ("guidebooks for mountain hiking in the summer season in Southern Norway") but by a different author. In this test 28 % of the sentences were translated, and 22 % with acceptable results. Finally, we have carried out evaluations using a syntactic test suite that cover most syntactic Norwegian constructions, while using a simple vocabulary from the training vocabulary lists. Here 88 % of the sentences were translated, out of which 63 % had good quality. This means that 55 % of all the syntactic test sentences were translated with good results.

To conclude, the LOGON demonstrator has reached a high level of translation quality for individual grammatical constructions, shown by the 55 % acceptably translated syntactic test sentences. But new vocabulary, even from the same domain, presents a major obstacle for the system, shown by the fact that only 22 % of the test texts from the same domain, but with different author, were acceptably translated.

4. References

Elliott, Debbie, Anthony Hartley and Eric Atwell (2004). A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. In R.E. Frederking and K.B. Taylor (Eds.): *AMTA. LNAI* 3265. Berlin: Springer-Verlag, pp. 67-73.

Halliday (T.C.) and Briss (E.A.) (1977). *The evaluation and systems analysis of the SYSTRAN machine translation system*.-NTIS, ADA 036.070 January 1977.

Hamon, Olivier, Anthony Hartley, Andrei Popescu-Belis and Khalid Choukri. 2007. Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA. In *Proceedings from MT Summit* Copenhagen.

Hovy, Eduard, Margaret King and Andrei Popescu-Belis. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, Volume 17, Issue 1, pp. 43--75

Johannessen, Janne Bondi, Torbjørn Nordgård and Lars Nygaard. Evaluation of translation output. To appear.

Koehn, Philipp. 2004a. PHARAOH - a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models (User Manual and Description for Version 1.2). Technical report, USC Information Sciences Institute.

Koehn, Philipp. 2004b. PHARAOH - Training Manual. Technical report, MIT. CSAIL.

Lie, Svein. (2003). *Innføring i norsk syntaks*. Oslo: Universitetsforlaget.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. New York*: IBM Research Report*.

Sparck Jones, K. and J.R.Galliers. (1996). *Evaluating Natural Language Processing Systems.* Berlin: Springer Verlag.

Tessiore, Lorenzo og Walter v. Hahn. (2000). Functional

Validation of a Machine Intrpretation System: Verbmobile. In Wahlster, W. (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer, pp. 611--634.

Oepen, Stephan, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation. On linguistics and probabilities in MT. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation,* Skövde, Sweden.

Van Slype, G. (1979). *Critical Methods for Evaluating the Quality of Machine Translation.* Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk.

Velldal, Erik. 2007. *Empirical Realization Ranking*. University of Oslo: PhD Thesis.

Velldal, Erik and Stephan Oepen. 2005. Maximum entropy models for realization ranking. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, pp. 109--116.

Test data

Hohle, Per. *Til fots i Jotunheimen [Hiking in Jotunheimen]*. Gyldendal, Oslo.

Lauritzen, Per Roger. *På tur i Jotunheimen / Huts and Hikes in Jotunheimen 1-4*. Cappelen, Oslo.

Unknown author. *Turglede [Tour pleasure]*. Trondhjems Turistforening, Trondheim.

Thomsen, Hanne and Kjell Helle-Olsen. *Turguide Preikestolen [Tour guide Pulpit Rock]*. Stavanger Turistforening, Stavanger.