Machine Translation as a testbed for multilingual analysis

Richard Campbell, Carmen Lozano, Jessie Pinkham and Martine Smets*

Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
{richcamp, clozano, jessiep, martines}@microsoft.com
*to whom all correspondence should be addressed

Abstract

We propose that machine translation (MT) is a useful application for evaluating and deriving the development of NL components, especially in a wide-coverage analysis system. Given the architecture of our MT system, which is a transfer system based on linguistic modules, correct analysis is expected to be a prerequisite for correct translation, suggesting a correlation between the two, given relatively mature transfer and generation components. We show through error analysis that there is indeed a strong correlation between the quality of the translated output and the subjectively determined goodness of the analysis. We use this correlation as a guide for development of a coordinated parallel analysis effort in 7 languages.

1 Introduction

The question of how to test natural language analysis systems has been central to all natural language work in the past two decades. It is a difficult question, for which researchers have found only partial answers. The most common answer is component testing, where the component is compared against a standard of goodness, usually the Penn Treebank for English (Marcus *et al*., 1993), allowing a numerical score of precision and recall (e.g. Collins, 1997).

Such methods have limitations, however, and need to be supplemented by additional methods. One limitation is the availability of annotated corpora, which do not exist for all languages. Secondly, comparison to an annotated corpus can only measure how well a system produces the kind of analysis for which the corpus is annotated, e.g. labeled bracketing of surface syntax. Evaluation of analysis of deeper, more semantically descriptive, levels requires additional annotated corpora, which may not exist. A more fundamental limitation of such methods is that they measure the goodness of a grammar without taking into account what the grammar is good for. This limitation is overcome, we claim, only by measuring the goodness of a grammar by its success in real-world applications.

We propose that machine translation (MT) is a good application to evaluate and drive the development of analysis components when the transfer component is based on linguistic modules. Multi-lingual applications such as MT allow evaluation of system components that overcomes the limitations mentioned above, and therefore serves as a useful complement to other evaluation techniques. Another significant advantage to using MT as a testbed for the analysis system is that it prioritizes analysis problems, highlighting those problems that have the greatest negative effect on translation output.

In this paper, we give an overview of NLPWin, a multi-application natural language analysis and generation system under development at Microsoft Research (Jensen *et al*., 1993; Gamon *et al*., 1997; Heidorn 2000), incorporating analysis systems for 7 languages (Chinese, English, French, German, Japanese, Korean and Spanish). Our discussion focuses on a description of the three components of the analysis system (called *sketch*, *portrait* and *logical form*) with a particular emphasis on the logical form derived as the end-product, which serves as the medium for transfer in our MT system.

We also give an overview of the architecture of the MSR-MT system, and of the evaluation we use to measure correctness of the translations. We demonstrate the correlation between the scores

assigned to translation outputs and the correctness of the analysis, using as illustration two language-pairs at different stages of development: Spanish-English (SE) translation, as a testbed for the Spanish analysis system, and French-English (FE) translation, as a testbed for the French analysis system.

2 Overview of the analysis component of NLPWin

Analysis produces three representations for the input sentence: sketch, portrait and logical form[1]. Sketch is the initial tree representation for the sentence, along with its associated attribute-value structure. An example of sketch is given in Figure 1, which shows the sketch tree for sentence (1).

(1)
Ce format est pris en charge par Windows 2000
this format is taken in charge by Windows 2000
'This format is supported by Windows 2000'

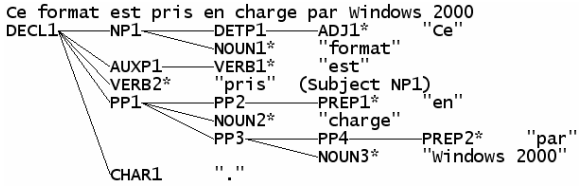

Figure 1 : Sketch analysis of (1)

Attachment sites for post-modifiers are not determined in sketch. In most cases, the information available as the syntactic tree is built is not sufficient to determine where e.g. prepositional phrases or relative clauses should be attached. Post-modifiers are thus systematically attached to the closest possible attachment site, and reattached, if necessary, by the reattachment module, a set of heuristic rules.

Reattachment rules apply to the sketch to produce the portrait; the portrait analysis of (1) is given in Figure 2, where the PP expressing the agent of the passive construction, originally attached to PP1 in sketch (see Figure 1) has been reattached at the sentence level.

[1] The presentation of the analysis module is very simplified, but sufficient for our current discussion. More details can be found in the references.

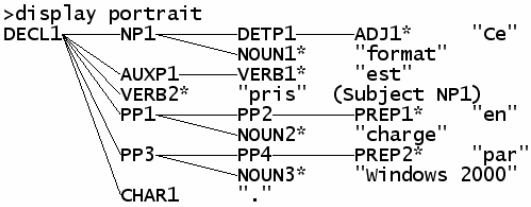

Figure 2: **Portrait analysis of (1)**

The portrait is the input to the computation of the logical form (LF), a labeled directed unordered graph representing the deep syntactic relations among the content words of the sentence (i.e., basic predicate-argument structure), along with some semantic information, such as functional relations expressed by certain prepositions.[2] At this level, the difference between active and passive constructions is normalized; control relations and long-distance dependencies, such as subjects of infinitives, arguments associated with gaps, etc., are resolved. The LF of (1) is shown in Figure 3. Note that the surface subject of the passive is rendered as the Dobj (deep object) in LF, and the *par*-phrase as the Dsub (deep subject).

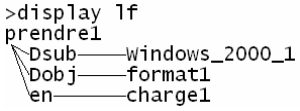

Figure 3 : LF analysis of (1)

Modifications to any of the analysis components are tested using monolingual regression files containing thousands of analyzed sentences; differences caused by the modification are examined manually by the linguist responsible for the change (Suzuki, 2002). This process serves as an initial screening to ensure that modifications to the analysis have the desired effect.

3 MSR-MT

In this section we review the basics of the MSR-MT translation system and its evaluation. The reader is referred to Pinkham *et al.* (2001) and Richardson *et al.* (2001) for further details on the French and Spanish versions of the system. The overall architecture and basic component structure

[2] LF as described here corresponds to the PAS representation of Campbell and Suzuki (2002).

are the same for both the FE and SE versions of the system.

3.1 Overview

MSR-MT uses the broad coverage analysis system described in Section 2, a large multi-purpose source-language dictionary, a learned bilingual dictionary, an application independent target-language generation component and a transfer component.

The transfer component consists of transfer patterns automatically acquired from sentence-aligned bilingual corpora (described below) using an alignment algorithm described in detail in Menezes and Richardson (2001). Training takes place on aligned sentences which have been analyzed by the source- and target-language analysis systems to yield logical forms. The logical form structures, when aligned, allow the extraction of lexical and structural translation correspondences which are stored for use at runtime in the transfer database. See Figure 4 for an overview of the training process.

The transfer database is trained on 350,000 pairs of aligned sentences from computer manuals for SE, and 500,000 pairs of aligned Canadian parliamentary data (the Hansard corpus) for FE.

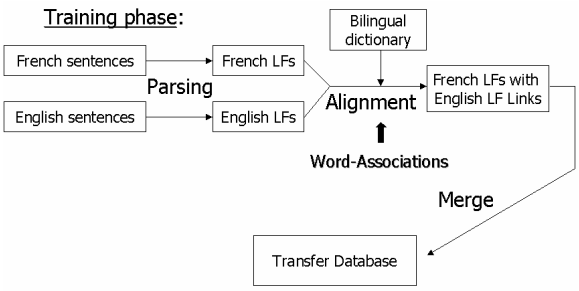

Figure 4: MSR-MT training phase

3.2 Evaluation of MSR-MT

Seven evaluators are asked to evaluate the same set of sentences. For each sentence, raters are presented with a reference sentence, the original English sentence from which the human French and Spanish translations were derived, and MSR-MT's machine translation.[3] In order to maintain consistency among raters who may have different levels of fluency in the source language, raters are not shown the original French or Spanish sentence (for similar methodologies, see Ringger *et al*., 2001; White *et al*., 1993).

All the raters enter scores reflecting the absolute quality of the translation as compared to the reference translation given. The overall score of a sentence is the average of the scores given by the seven raters. Scores range from 1 to 4, with 1 meaning **unacceptable** (not comprehensible), 2 meaning **possibly acceptable** (some information is transferred accurately), 3 meaning **acceptable** (not perfect, but accurate transfer of all important information, and 4 meaning **ideal** (grammatically correct and all the important information is transferred).

4 Examples from FE and SE

In this section we discuss specific examples to illustrate how results from MT evaluation help us to test and develop the analysis system.

4.1 FE translation: the Hansard corpus

The evaluation we are discussing in this section was performed in January 2002, at the beginning of our effort on the Hansard corpus. The evaluation was performed on a corpus of 250 sentences, of which 55.6% (139 sentences) were assigned a score of 2 or lower, 30.4% (76 sentences) were assigned a score greater than 2 but not greater than 3, and 14% (35 sentences) were assigned a score greater than 3.

Examination of French sentences receiving low-score translations led to the identification of some classes of analysis problems, such as the following:

- mis-identification of vocatives
- clefts not represented correctly
- mis-analysis of *ce qui* / *ce que* free relatives
- bad representation of complex inversion (pronoun-doubling of inverted subject)
- no treatment of reflexives
- fitted parses (i.e., not spanning the sentence)

Most of the problematic structures are characteristic of spoken language as opposed to more formal, written styles (vocatives, clefts, direct questions), and had not been encountered in our previous work, which had involved mostly translation of technical manuals. Other problems

[3] Microsoft manuals are written in English and translated by hand into other languages. We use these translations as input to our system, and translate them back into English.

(free relatives, reflexives) are analysis issues that we had not yet addressed. Fitted parses are parses that do not span the whole sentence, but are pieced together by the parser from partial parses; fitted parses usually result in poor translations.

Examples of translations together with their score are given in Table I. The source sentences are the French sentences, the reference sentence is the human translation to which the translation is compared by the evaluators, and the translation is the output of MSR-MT. Each of the three categories considered above is illustrated by an example.

Sentence (2) (with a score of 1.5) is a direct question with complex inversion and the doubled subject typical of that construction. In the LF for (2), *les ministres des finances* is analyzed as a modifier, because the verb *réunir* already has a subject, the pronoun *ils* 'they'. There are a couple of additional problems with this sentence: *si* is analyzed as the adverb meaning 'so' instead of as the conjunction meaning 'if', and a direct question is analyzed as a complement clause; the sketch and LF analyses of this sentence are given in the Appendix.. The MSR-MT translation of this sentence has a very low score, reflecting the severity of the analysis problems.

The two other sentences, on the other hand, do not have analysis problems: the poor translation of (3) (score 2.16) is caused by bad alignment (*droit* translates as *right* instead of *law*), and the translation of (4) (score 3) is not completely fluent, but this is due to an English generation problem, rather than to a French analysis problem. This last sentence is the most correct with appropriate lexical items and has the highest score of the three.

Of the 139 sentences with score 2 or lower, 73% were due to analysis problems, and 24% to alignment problems. Most of the rest had bugs related to the learned dictionary. There were a few cases of very free translations, where the reference translation was very far from the French sentence, and our translation, based on the source sentence, was therefore penalized.

These figures show that, at this stage of development of our system, most of the problems in translation come from analysis. Translation can be improved by tackling analysis problems exhibited by the lowest scoring sentences, and, conversely, analysis issues can be discovered by looking at the sentences with the lowest translation score.

The next section gives examples of issues with the SE system, which is more mature than the FE system.

4.2 SE translation: Technical manuals

An evaluation of the Spanish-English MT system was also performed in January 2002, after work on the MT system had been progressing for approximately a year and a half. The SE system was developed and tested using a corpus of sentences from Microsoft technical manuals. A set of 600 unseen sentences was used for the evaluation.

Out of a total of 600 sentences, the number of sentences with a score from 3 to 4 was 251 (42%), the number of sentences with a score greater than 2 but less than 3 was 186 (31%), and the remaining 163 sentences, (27%) had a score of 2 or lower. Of these 163 sentences with the lowest scores, 50% (82 sentences) had analysis problems, and 17% of them (29 sentences) had fitted parses. A few of the fitted parses, 7 sentences out of 29, had faulty input, e.g. input that contained unusual characters or punctuation, typos, or sentence fragments.

Typical analysis problems that led to poor translation in the SE system include the following:

- incorrect analysis of arguments in relative clauses, especially those with a single argument (and a possible non-overt subject)
- failure to identify the referent of clitic *le* (i.e. *usted* 'you') in imperative sentences in LF
- mis-analysis of Spanish reflexive or *se* constructions in LF
- incorrect syntactic analysis of homographs
- incorrect analysis of coordination
- mis-identification of non-overt or controlled subjects
- fitted parses

Table II contains sample sentences from the SE evaluation. For each row, the second column displays the Spanish source sentence with the reference sentence in the next column, the translation produced by the MT system is in the fourth column, and the score for the translation assigned by the human evaluators in the last column.

#	Source	Reference	Translation	Score
(2)	Si tel n'était pas le cas, pourquoi les ministres des Finances des provinces se seraient-ils réunis hier pour essayer de s'entendre sur un programme commun à soumettre au ministre des Finances?	If that were not the case, why were the finance ministers of the provinces coalescing yesterday to try and come up with a joint program to bring to the finance minister?.	Not was the case that they have the ministers met why yesterday Finances of the provinces trying to agree on a common program to bring Finances for the minister this so like?	1.5
(3)	Nous ne pouvons pas appuyer cette motion après que le Bloc québécois ait refusé de reconnaître la primauté du droit et de la démocratie pour tous.	We cannot support this motion after seeing the Bloc Quebecois refuse to recognize the rule of law and the principle of democracy for all.	We cannot support this motion after the Bloc Quebecois has refused to recognize the rule of the right and democracy for all.	2.16
(4)	En tant que membre de l'opposition officielle, je continuerai d'exercer des pressions sur le gouvernement pour qu'il tienne ses promesses à cet égard.	As a member of the official opposition I will continue to pressure the government to fulfil its promises in this regard.	As member of the official opposition, I will continue to exercise pressures on the government for it to keep its promises in this regard.	3

Table I: Examples of FE translation

#	Source	Reference	Translation	Score
(5)	Este procedimiento sólo es aplicable si está ejecutando una versión de idioma de Windows 2000 que no coincida con el idioma en el que desee escribir.	This procedure applies only if you are running a language version of Windows 2000 that doesn't match the language you want to type	This procedure only applies if you are running a Windows 2000 language version that does not match the language that you want to type.	3.8
(6)	Repita este proceso hasta que haya eliminado todos los componentes de red desde las propiedades de Red, haga clic en Aceptar y, a continuación, haga clic en Sí cuando se le pregunte si desea reiniciar el equipo.	Repeat this process until you have deleted all of the network components from Network properties, click OK, and then click Yes when you are prompted to restart your computer.	Repeat this process until you have deleted all of the network components from the Network properties, you click OK, and you click Yes then when asking that to restart the computer is wanted for him.	2.0
(7)	En el siguiente ejemplo se muestra el nombre de la presentación que se está ejecutando en la ventana de presentación con diapositivas uno.	The following example displays the name of the presentation that's currently running in slide show window one.	In the following example, the display name that is being run in the slide show window is displayed I join.	1.4

Table II: Examples of SE translation

In the evaluation process, human evaluators compared the MT translation to the reference sentence, in the manner described in Section 4.1.

Example (5), with a score of 3.8, illustrates the fact that human evaluators considered the translation 'a Windows 2000 language version' to be a slightly worse translation than 'a language version of Windows 2000' for *una version de idioma de Windows 2000*; however the difference is so slight as to not be considered an analysis problem.

Example (6) illustrates the failure to identify *usted* 'you' (understood as the subject of the imperative) as the referent of the pronominal clitic *le*; as mentioned above, this is a common source of bad SE translations. The last example (7) is a sentence with a fitted parse due to misanalysis of a word as its homograph : *uno* is analyzed as the first person singular present form of the verb *unir* 'join' instead of as the noun *uno* 'one'; the LF of this sentence is given in the Appendix.

4.3 Discussion

The examples discussed in this section are typical: The sentences for which MSR-MT produces better translations tend to be the ones with fewer analysis

errors, while those which are misanalyzed tend to be mistranslated.

In this way, evaluation of MT output serves as one way to prioritize analysis problems; that is, to decide which among the many different analysis problems lead to the most serious problems. For example, the poor quality of the translation of (2) highlights the need for an improved analysis of complex inversion in the French grammar, which will need to be incorporated into the sketch and/or LF components. Similarly, the poor translation of (7) indicates the need to deal better with homographs in the Spanish morphological or sketch component.

More generally, the analysis of FE and SE translation problems has led to the lists of analysis problems given in Sections 4.1 and 4.2, respectively. Analysis problems identified in this way then become priorities for grammar/LF development.

5 Conclusion

We have outlined how the output of MT can be used as testbed for linguistic analysis in the source language, supplementing other methods. The main advantage of this approach, in our view, is that it helps to prioritize analysis problems, highlighting those which have the most direct bearing on the application(s), the correct functioning of which is the main goal of the system.

Acknowledgements

This paper represents the work of many people in the NLP group at MSR; we acknowledge their contributions.

References

Campbell, R. and H. Suzuki. 2002. Language-neutral representation of syntactic structure. In R. Malaka, R. Porzel and M. Stube, eds., Proceedings of the First International Workshop on Scalable Natural Language Understanding.

Collins, M. 1997. Three generative, lexicalised models for statistical parsing. Proceedings of the 35th Annual Meeting of the ACL, Madrid.

Gamon, M., C. Lozano, J. Pinkham and T. Reutter. 1997. Practical experience with grammar sharing in multilingual NLP. In Burstein J., Leacock C., eds, Proceedings of the Workshop on Making NLP Work, ACL Conference, Madrid.

Heidorn, G. 2000. Intelligent writing assistance. In R. Dale, H. Moisl and H. Somers, eds., Handbook of Natural Language Processing.

Jensen, K., G. Heidorn and S. Richardson, eds. 1993. Natural Language Processing: The PLNLP Approach, Boston, Kluwer.

Marcus, M., B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. In Proceedings of the 31st Annual Meeting of the ACL.

Menezes, A. and S. Richardson. 2001. A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In Proceedings of the Data-Driven MT workshop, ACL 2001.

Pinkham, J., M. Corston-Oliver, M. Smets and M. Pettenaro, 2001. Rapid assembly of a large-scale French-English MT system. In Proceedings of the 2001 MT Summit.

Richardson, S., W.B. Dolan, A. Menezes and J. Pinkham. 2001. Achieving commercial-quality translation with example-based methods. In Proceedings of the 2001 MT Summit.

Ringger, E.K., M. Corston-Oliver, and R.C. Moore. 2001. Using Word-Perplexity for Automatic Evaluation of Machine Translation. Unpublished ms.

Suzuki, H. 2002. A development environment for large-scale multi-lingual parsing systems. Workshop on Grammar Engineering and Evaluation, COLING 2002.

White, J.S., T.A. O'Connell, and L.M. Carlson. 1993. Evaluation of machine translation. In Human Language Technology: Proceedings of a Workshop (ARPA). 206-210.

Appendix

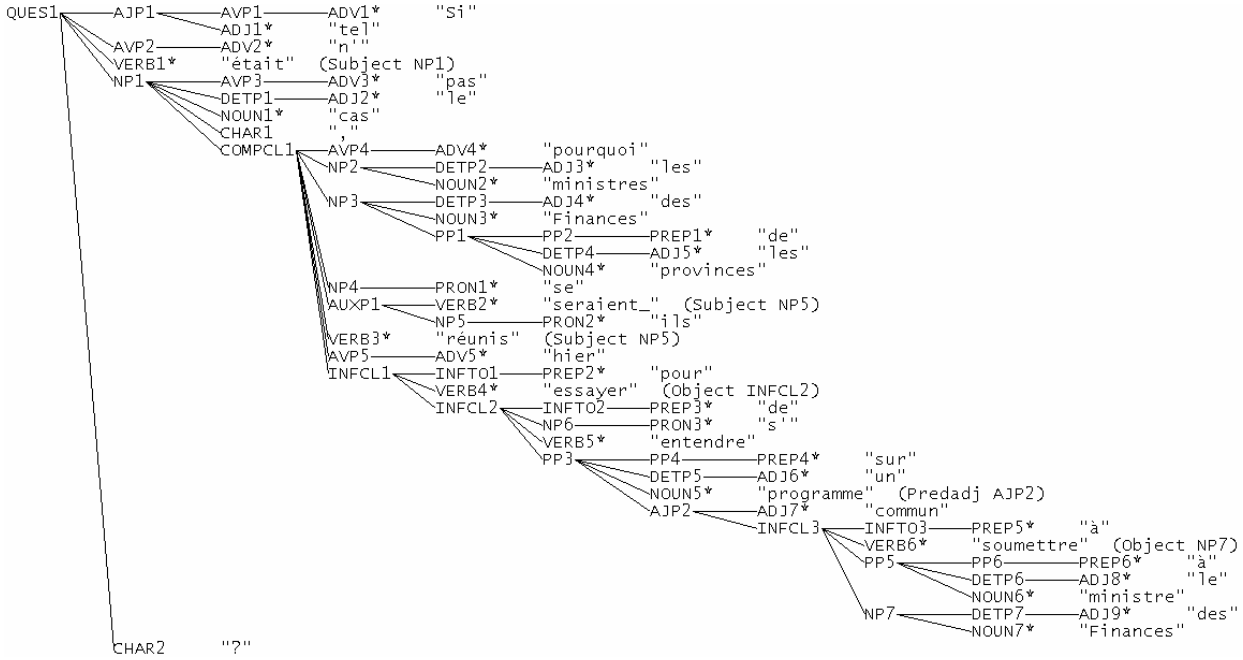

Figure 5 : Sketch analysis of (2)

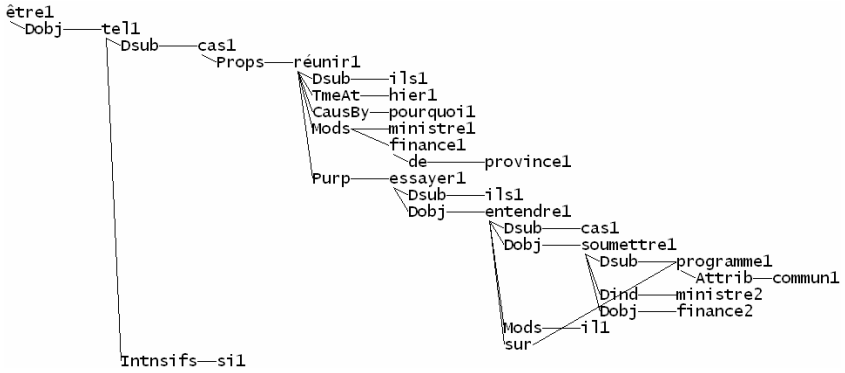

Figure 6 : LF analysis of (2)

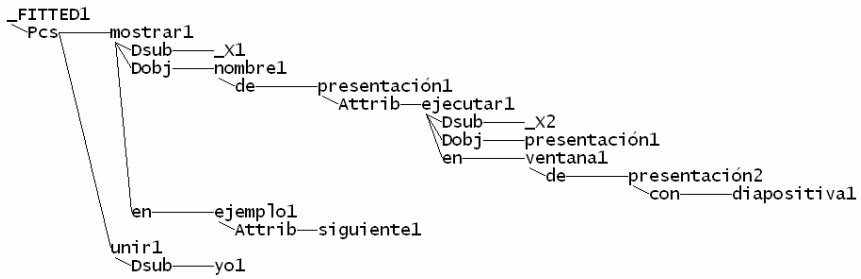

Figure 7 : LF analysis of (7)