Round-Trip Translation: What Is It Good For?

Harold Somers
School of Informatics
Manchester University
Manchester, UK
Harold.Somers@manchester.ac.uk

Abstract

This paper considers the popular but questionable technique of 'round-trip translation' (RTT) as a means of evaluating free on-line Machine Translation systems. Two experiments are reported, both relating to common requirements of lay-users of MT on the web. In the first we see whether RTT can accurately predict the overall quality of the MT system. In the second, we ask whether RTT can predict the translatability of a given text. In both cases, we find RTT to be a poor predictor of quality, with high BLEU and F-scores for RTTs when the forward translation was poor. We discuss why this is the case, and conclude that, even if it seemed obvious that RTT was good for nothing, at least we now have some tangible evidence.

1 Introduction

Macklovitch (2001:27) talks of "the spectacular growth and pervasiveness of the World Wide Web" leading to a "democratization" of Machine Translation (MT) which has "profoundly transformed the MT business". The availability of free on-line MT systems, since CompuServe's initial experiments in 1994 (Flanagan, 1996) and then more significantly AltaVista's collaboration with Systran from 1997 onwards (Yang and Lange, 1998), has indeed revolutionized the MT world, creating a whole new and significantly large community of users, mostly with little or no knowledge or understanding of how MT works or even, in some cases, how *language* works. Such users are, nevertheless, keen to know how good MT output is, and frequently resort to the intuitive technique of 'round-trip translation' (RTT), or 'back-and-forth translation', in which they take a given text or sentence, have it translated into some foreign language by the MT system (the 'forward translation', henceforth FT), then have it translated back into the original language by the same system (the 'back translation', BT).

Popular articles on MT by journalists and other lay-users all too frequently use this technique to 'evaluate' MT, with results which are, depending on your predisposition, hilarious or infuriating. A recent example is from the *Biomedical Translations* website (Anon, 2003), where the author explains the technique, and suggests that "In theory, the back translated English should match the original English." Several garbled examples are then given, and the article concludes "Would you trust your surgeon using these instructions?" Another website recognizes the problem "Machine translations can produce text that is garbled or hilariously inaccurate", and suggests as a resolution "Test the precision of your translated text by sending a phrase on a round trip through the translation engine." (Anon, 2005).

The dangers of this approach have long been appreciated: for example Huang (1990), addressing the problem of evaluating output when you do not know the target language, describes it as the "seemingly most natural way" to evaluate a translation, but quickly warns that the results are not reliable. More recently, O'Connell (2001) gives the following sound advice on an IBM website:

> "A common misunderstanding about MT evaluation is the belief that back translation can disclose a system's usability. […] The theory is that if back translation returns [the source language] input exactly, the system performs well for this language pair. In reality, evaluators cannot tell if errors occurred during the passage to [the target language] or during the return passage to [the source language]. In addition,

Proceedings of the Australasian Language Technology Workshop 2005, pages 127–133,
Sydney, Australia, December 2005.

> any errors that occur in the first translation […] cause more problems in the back translation."

So, although it is widely agreed in the MT community that RTT is a bad technique, and equally widely suggested in the lay community that it is an effective way to evaluate systems, there has been little or no work to demonstrate empirically whether RTT is in fact as misleading as it is claimed.

In the next section we will briefly review the reasons why one might be wary of RTT as an indicator of MT quality: while one can cite anecdotal evidence of bad round trips, we can ask whether on a larger scale RTT might after all be indicative at least of general trends. In this regard, given the situation that lay users finds themselves in, we will consider two issues of concern to them: Which is the best MT system to use? And how machine-translatable is my text? In Sections 3 and 4 we will present two experiments that take this user's need, and explore whether RTT can meet it, or not.

2 Two perspectives on RTT

2.1 Why RTT might not work

As O'Connell (2001) states in the earlier quote, and as other commentators have pointed out, RTT could be misleading for three reasons:

First, if the round trip is bad, you cannot tell whether it was the outward journey or the return trip where things went wrong. For example, (1) shows an RTT from English to Italian and back again using Babelfish. The resulting BT (1c) is garbled, but in fact apart from a possible gender error (loan words usually take the gender of their literal translation, so *Home Page* should probably be feminine) the forward translation into Italian is really quite acceptable.

(1) a. Select this link to look at our home page.
b. *Selezioni questo collegamento per guardare il nostro Home Page.*
c. Selections this connection in order to watch our Home Page.

Of course, if it is the outward journey that is bad, then the return trip could be bad, but it might be disproportionately so, because of the old maxim 'garbage in garbage out'.

However, and this is the second point, a bad FT can nevertheless lead to a quite reasonable BT. So the fact that the round trip gives a good result does not necessarily tell you anything about the outward journey. This can be illustrated in (2), again using Babelfish, where the idiomatic phrase is translated literally into meaningless Portuguese (2a) and then 'perfectly' back into Enlgish (2c).

(2) a. tit for tat
b. *melharuco para o tat*
c. tit for tat

The third point is that of course the basic premise of RTT is flawed: even a pair of human translators would not be expected to complete a perfect RTT, in the sense that the return translation would be word-for-word identical to the original source text.

2.2 Why RTT might appear to work

So, it is easy to show RTT not working. But equally we should acknowledge that sometimes, RTT *does* appear to work, producing a quite understandable paraphrase and, if only we knew it, a reasonable translation on the way. Examples (3) and (4), translated by Freetranslation, illustrate this.

(3) a. The spirit is willing but the flesh is weak.
b. *Дух желает, но плоть слаба.*
c. The spirit wishes, but the flesh is weak.

(4) a. Once, a jolly swagman camped beside a billabong.
b. *Однажды, веселая бродяга разбила лагерь около устья реки.*
c. Once, the cheerful tramp has broken camp about a mouth of the river.

The conclusion is that *for a single given sentence*, we cannot know for sure if a good (or bad) RTT indicates that the FT was good (or bad) or vice versa. But it is a not unreasonable hypothesis that over the length of a longer text, average RTT quality might reflect the general quality of the system used. This will be the subject of our first experiment, reported in Section 3.

2.3 Lay-users and MT

For some time now, observers of MT have identified two distinct uses of MT, labelled 'for assimilation' and 'for dissemination'. The differences between the two are neatly summarized in Table 1. Until now, use of free on-line web-based MT services has been assumed to lie more or less firmly on the 'assimilation' side. However, as the availability of the service has become well known, we now see a lot of web pages with explicit links to

Assimilation	Dissemination
many SLs, one TL	one SL, many TLs
any style	controlled style
any topic	restricted topic(s)
poor quality OK	good quality required
post-editing if needed	no post-editing
user is reader	user is author

Table 1. Differences between MT for assimilation and MT for dissemination

MT services, which means that web-page designers now see on-line MT as a way of getting their message translated. They have become users of MT for dissemination, although typically they do not fit the profile outlined in Table 1, and it has been argued (Gaspari, 2004; Somers and Gaspari, 2005) that web-page designers need to be better educated about what MT can and cannot do.

It is not unreasonable therefore for web-page designers to seek some way of knowing how well their web pages will be translated well by free on-line MT services. There has been a fair amount of research recently on 'translatability'. (Gdaniec, 1994; Bernth, 1999a,b; Bernth and McCord, 2000; Underwood and Jongejan, 2001; Bernth and Gdaniec, 2002; O'Brien, in press). Research has focussed on identifying 'translatability indicators', stylistic or grammatical linguistic features that are known to be problematic for MT (so a more transparent name would perhaps be 'translation difficulty indicators'). For example, mid-sentence parenthetical statements or the use of the passive voice could respectively be classified as stylistic and grammatical indicators. While such measures are of use to linguists, and to designers of controlled languages, they mean little or nothing to the average lay-user. Even if RTT is not reliable on an individual sentence-by-sentence basis, might it be reliable enough to show whether an entire document is *by and large* machine translatable?

3 Experiment 1: Can RTT tell us which system is best?

Even if RTT does not always work, we might hope that the quality of the RTT will reflect the quality of the FT: if this is true, then at least RTT could be used to help lay-users to decide which system to use, when they are faced with a large number to choose from. In order to explore this hypothesis, we set up a first experiment in which we took four texts representing various language pairs, translated them each using five free on-line MT systems,[1] then translated the resulting FT back into the original language using the same system. We used two standard measures to evaluate the results, the familiar BLEU metric (Papineni et al., 2002), and Turian et al.'s (2003) F-score metric. A number of researchers have commented on the fact that BLEU scores do not always agree with the F-score, based on precision and recall, nor sometimes with human judgments, especially for shorter stretches of text (e.g. Way and Gough, in press). In addition, alternative packages available on the Web offering implementations of BLEU give significantly different results. Accordingly, we used our own implementations of BLEU and F-score,[2] and show results with both metrics; while they sometimes rank the translations differently, they both tell the same overall 'story', as we shall see.

The texts were as follows: extracts from the French web pages of the Tourist Offices of Marseilles and Barèges (a skiing resort) for translation into English, and two passages from the Europarl corpus of European Parliament Proceedings 1996–2003, one in English, for translation into German, and one in French, for translation into English. All the texts were around 100 sentences long.

Figures 1 and 2 show the BLEU and F-scores for the 20 pairs of translations, FT and BT, grouped by text, and ordered within each group. The order of the systems is different for each text, but since, for our purposes, we are only interested in seeing whether the scores for the FTs and BTs correlate, the identity of the individual systems is unimportant. The first thing to notice is that both BLEU and F-score show little correlation between the FT and BT scores. Figure 3 shows this more strikingly, as does a Pearson's coefficient of $r = -0.04$ for these scores. The F-scores (not shown here) correlate somewhat better, at $r = 0.61$, but this is nowhere near useful for our purposes.

However, we should note that, as Figures 1 and 2 show, the difference in scores for some of these systems is really quite small. Also, for two of the texts, the system with the top-ranking score for FT is actually ranked 4th or 5th for the BT.

[1] Babelfish, Freetranslation, Systran, ProMT, and Worldlingo.

[2] Thanks to Simon Zwaarts for these, and also for Python scripts used to translate on-line large amounts of material, overcoming the text-length limits imposed by many of the systems.

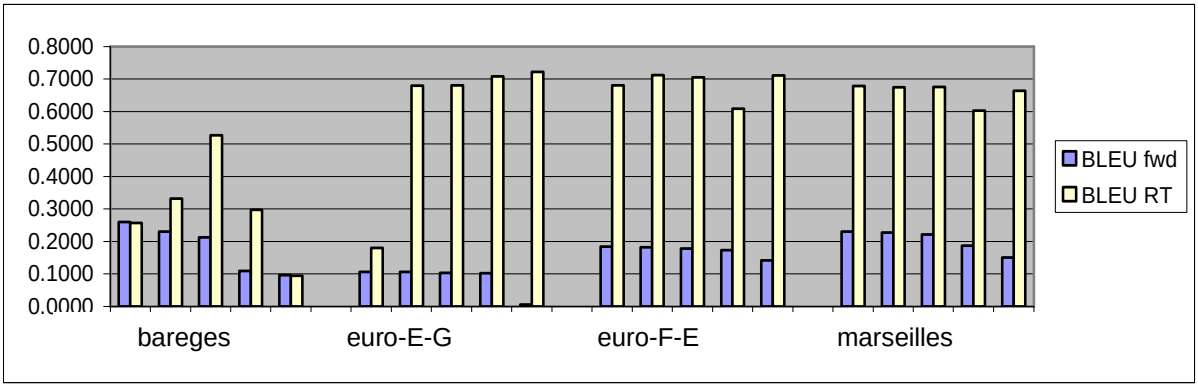

Figure 1. BLEU scores for the 20 forward and round-trip translations.

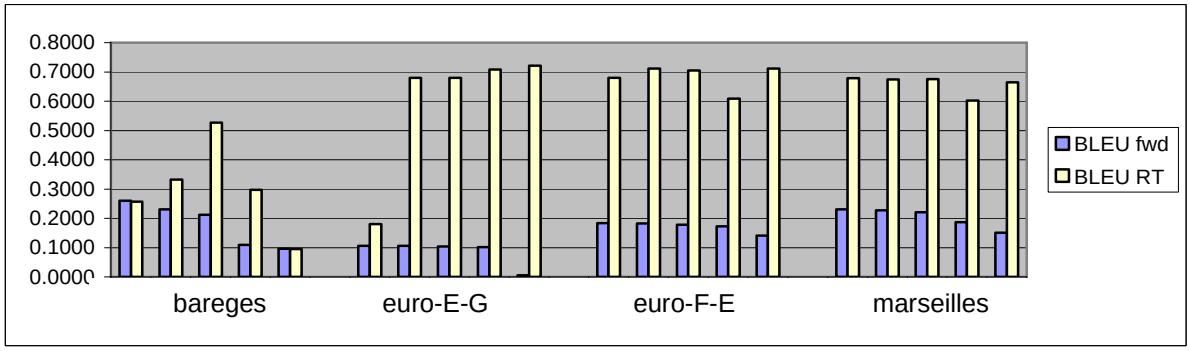

Figure 2. F-scores for the 20 forward and round-trip translations.

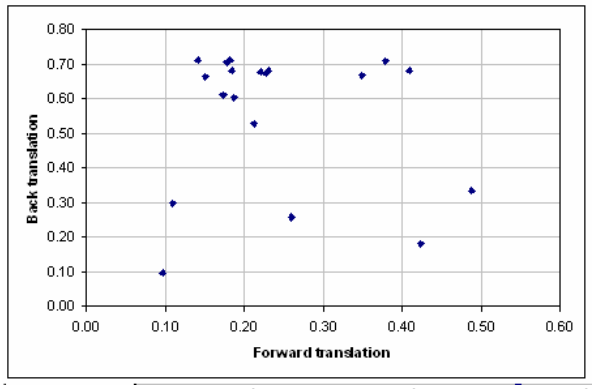

Figure 3. Correlation of BLEU scores for FT and BT. If the scores correlated well, they would cluster around the diagonal.

Our first conclusion then is that RTT is *not* a particularly good way to identify which system is better: if anything, a high-scoring BT indicates either the best or the worst system, but even this is not systematic.

What is also striking is that the BT score is often better than the FT score, and the difference is greatest when the FT score is low. Although the results do not show a consistent pattern, what is clear is that a good score for the BT generally does not necessarily 'predict' a good score for the FT; rather more often the opposite.

The reason for this is fairly easy to explain, considering how these MT systems in general work. Although systems perform source-text analysis to a certain extent, when all else fails they resort to word-for-word translation, and where there is a choice of target word they will go for the most general translation. Clearly, when the input to the process is difficult to analyse, the word-for-word translation will deliver pretty much the same words in the BT as featured in the original text. A further point to make is that the BLEU metric (and to a lesser extent the F-score) 'reward' word matches, even if the word order is somewhat scrambled. This can be illustrated with (5) which shows the source text (5a) and model translation (5b), the FT (5c) and the BT (5d). About two thirds of the words in (5a) appear in (5d); this pair of sentences alone would merit a BLEU score of 0.5882.

(5) a. All of us here are pleased that the courts have acquitted him and made it clear that in Russia, too, access to environmental information is a constitutional right.

b. *Wir freuen uns hier alle, daß das Gericht ihn freigesprochen und deutlich gemacht hat, daß auch in Rußland der Zugang zu Umweltinformationen konstitutionelles Recht ist.*

c. *Alle, von den uns hier erfreut werden, dass die Gerichtshöfe ihn freigelassen haben, und hat es reinigt gemacht, daß in Russland auch auf zu Umweltinformationen zugreift ist ein verfassungsmäßiges Recht.*

d. Everyone of which here delighted become us that the courts it released have has cleans made, and it that in Russia also on to environment information take action, is a constitutional right.

4 Experiment 2: Can RTT tell us how well our text will be translated?

In our second experiment, we wanted to see if the scores for the RTT would correlate with scores for FT when we compare texts that the MT systems translate well with texts that prove difficult. Based on the BLEU and F-scores, we took three of the texts from the first experiment, and the scores for their RTTs using Freetranslation, neither the best nor the worst of the MT systems. These 'hard' texts were the Marseilles web-page and the two Europarl examples. Against these we constructed three 'easy' texts: a children's story (*Goldilocks and the Three Bears*), some text from Canadian weather forecasts, and some typical entries from a tourist's phrase-book. We constructed the parallel *Three Bears* text and the tourist phrases from various websites. The weather bulletins come from RALI's Météo website;[3] we 'helped' the MT system by pre-editing the texts, converting the all-uppercase text to mixed case, inserting accents, and also changing *moins* and *minus* in temperature read-outs to a minus-sign. Like the 'hard' texts, the 'easy'texts were all roughly 100 lines long.

In (6) we see some examples of BTs that show that the easy texts were indeed generally well translated back and forth.

(6) a. Therefore she went in top in the bedroom where the three Bears slept, and there was the three beds.

b. Today. Cloudy with the clear periods and some snow. High close to –9. The winds of the west 15 to 30 km/h. Tonight. Cloudy with the clear periods and 30% probability of flurries.

c. Do you speak the English ? I do not speak the French. I do not understand. Please to speak slowly. I hope that you understand my English

The comparison of the BLEU scores for the FT and BT of these six texts is shown in Figure 4. The figure shows quite dramatically that, at least as far as the BLEU scores go, the easy texts are somewhat easier to translate than the hard texts; and it shows equally clearly that the score for the RTT does not reflect this at all: in fact according to the RTT score, all the texts are of about the same difficulty. The correlation between BLEU scores for the FT and BT is $r = -0.31$, while for the F-scores it is $r = 0.59$.

5 Conclusions

In this paper we have tried to demonstrate explicitly what most MT researchers already assumed: to paraphrase the words of Edwin Starr's 1970s anti-war song suggested by the title of this paper, "RTT (grunt), what is it good for? Absolutely nothing (say it again)...". This may have seemed like an obvious result, but we would like to restate that until now no one as far as we know has published results demonstrating this.

Before we leave the topic however, it may be appropriate to cast one small shadow of doubt over the result: throughout this work we have relied on the BLEU and F-score metrics to judge the translations. So our conclusion is really that RTT cannot tell good MT systems from bad ones, or easy-to-translate texts from hard ones, *based on automatic evaluation methods*. To be really sure of our results, we should like to replicate the experiments evaluating the translations using a more old-fashioned method involving human ratings of intelligibility. The reason for this is that both these metrics reward translations that are lexically close to the oracle translation, without taking into account whether they are grammatical or make sense. If we look at our high-scoring BTs, we can see that often they do indeed match the vocabulary of the model translations, without making much sense: contrast examples (5) and (6), both high scoring, but differing in qualities of grammaticality and intelligibil-

[3] http://rali.iro.umontreal.ca/meteo, as described in Langlais et al. (in press).

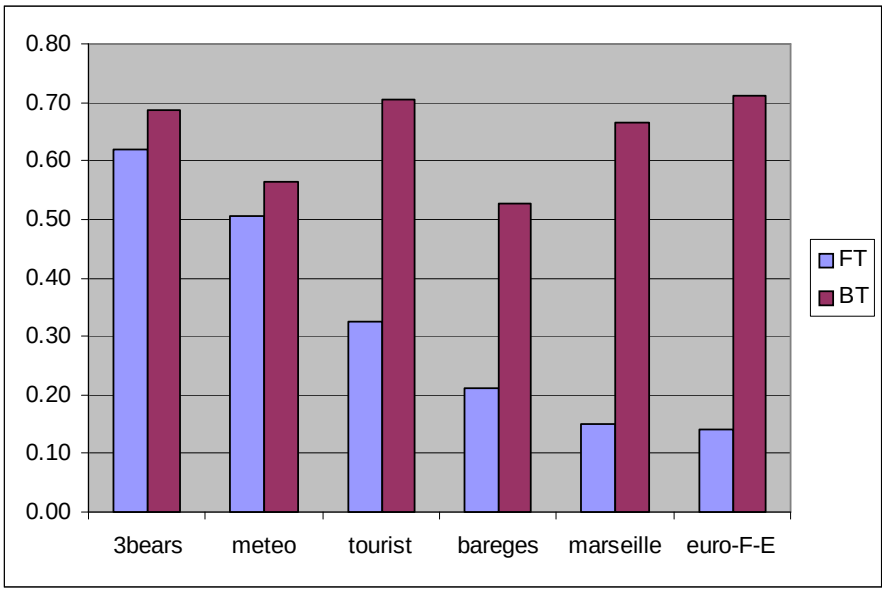

Figure 4. BLEU scores for the forward and back translations of three 'easy' and three 'hard' texts.

ity. So perhaps this is not quite the end of the story after all.

Acknowledgments

This work was completed while the author was on study leave at the Centre for Language Technology, Macquarie University. He is most grateful to all his colleagues there for their warm welcome, and especially Simon Zwaarts and Menno van Zaanen, whose programs and scripts were used to translate on-line large amounts of material, overcoming the text-length limits imposed by many of the systems, and to implement the BLEU and F-score measures. Thanks also to Elena Akhmanova for advice on the Russian examples in this paper, and to Steve Cassidy for suggesting the experiments reported here.

References

Anon. 2003. More Machine Translation: Fun with computer generated translation!, *Biomedical Translations*, News, October 2003. www.biomedical.com/news.html.

Anon. 2005, Gotcha!: Translation software. Software that translates text from one language to another may be a big help—or hindrance—to businesses and relief agencies alike. *Baseline*, May 2, 2005. www.baselinemag.com/article2/0,1397,1791588,00.asp.

Bernth, A. 1999a. EasyEnglish: A confidence index for MT. *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI '99*, Chester, England, pp. 120–127.

Bernth, A. 1999b. Controlling input and output of MT for greater user acceptance. *Translating and the Computer 21*, London, [pages not numbered].

Bernth, A. and C. Gdaniec. 2002. MTranslatability. *Machine Translation* **16**:175–218.

Bernth, A. and M. McCord. 2000. The effect of source analysis on translation confidence. In J.S. White (ed.) *Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico, ...*, Berlin: Springer, pp. 89–99.

Flanagan, M. 1996. Two years online: experiences, challenges and trends. *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Canada, 192–197.

Gaspari, F. 2004. Integrating on-line MT services into monolingual web-sites for dissemination purposes: An evaluation perspective. *9th EAMT Workshop Broadening Horizons of Machine Translation and its Applications*, Valletta, Malta, pp. 62–72.

Gdaniec, C. 1994. The Logos translatability index. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, pp. 97–105.

Huang, X. 1990. A Machine Translation system for the target language inexpert. *Papers presented to the 13th International Conference on Computational*

Linguistics, COLING-90, Vol. 3, Helsinki, pp. 364–367.

Langlais, P., S. Gandrabur, T. Leplus and G. Lapalme. In press. The long-term forecast for weather bulletin translation. To appear in *Machine Translation.*

Macklovitch, E. 2001. Recent trends in translation technology. *Proceedings of the 2nd International Conference, The Translation Industry Today: Multilingual Documentation, Technology, Market*, Bologna, Italy, pp. 23–47.

O'Brien, S. In press. Methodologies for measuring the correlations between post-editing effort and machine translatability. To appear in *Machine Translation*.

O'Connell, T.A. 2001. Preparing your web site for machine translation: how to avoid losing (or gaining) something in the translation. IBM website, www-128.ibm.com/developerworks/web/library/us-mt/.

Papineni, K., S. Roukos, T. Ward and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *ACL-02: 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pp. 311–318.

Somers, H. and F. Gaspari. 2005. The impact of free web-based Machine Translation services on internationalisation. Paper presented at APWSI 2005: The Asia Pacific Workshop on Software Internationalisation, at ASWEC 2005, The Australasian Software Engineering Conference, Brisbane, Qld.

Turian, J.P., L. Shen, and I.D. Melamed. 2003. Evaluation of machine translation and its evaluation. *MT Summit IX: Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, pp. 23–28.

Underwood, N. and B. Jongejan. 2001. Translatability checker: A tool to help decide whether to use MT. In *Proceedings of MT Summit VIII: Machine Translation in the Information Age*, Santiago de Compostela, Spain, pp. 363–368.

Way, A. and N. Gough. In press. Controlled translation in an example-based environment: What do automatic evaluation metrics tell us? To appear in *Machine Translation*.

Yang, J. and Lange, E.D. 1998. Systran on AltaVista: a user study on real-time machine translation on the Internet. In: D. Farwell, L. Gerber and E. Hovy (eds) *Machine Translation and the Information Soup*: *Third Conference of the Association for Machine Translation in the Americas, AMTA'98, Langhorne, PA,*Berlin: Springer, 275–285.