Detecting Inappropriate Use of Free Online Machine Translation by Language Students – A Special Case of Plagiarism Detection

Harold Somers
School of Informatics
University of Manchester
PO Box 88
Manchester M60 1QD,
UK

Federico Gaspari
School of Informatics
University of Manchester
PO Box 88
Manchester M60 1QD,
UK

Ana Niño
School of Education
University of Manchester
Oxford Road
Manchester M13 9PL, UK

Harold.Somers@manchester.ac.uk,
{F.Gaspari,A.Nino}@postgrad.manchester.ac.uk

Abstract. The ready availability of free online machine translation (MT) systems has given rise to a problem in the world of language teaching in that students – especially weaker ones – use free online MT to do their translation homework. Apart from the pedagogic implications, one question of interest is whether we can devise any techniques for automatically detecting such use. This paper reports an experiment which aims to address this particular problem, using methods from the broader world of computational stylometry, plagiarism detection, text reuse, and MT evaluation. A pilot experiment comparing 'honest' and 'derived' translations produced by 25 intermediate learners of Spanish, Italian and German is reported.

1. Introduction

One of the most important developments in the history of Machine Translation (MT) has been the availability, since about 1994, of free MT online: while initially perhaps a marketing ploy, this has had a profound effect on the perceptions of the general public, as well as shaping the development of the technology. It was CompuServe who first entered into an agreement with Systran to make MT available free online (Flanagan, 1996), though AltaVista's subsequent development of the Babelfish website is much better known. Some ten years on, numerous sites offer MT between vast numbers of language pairs, although some of them are little more than on-line dictionaries, sometimes of dubious quality. There have been a number of studies of the use of free online MT (FOMT) systems, both from the developers' and users' perspectives (Bennett, 1996; Miyazawa et al., 1999; Yang and Lange 1998, 2003). This paper concerns one small group of such users, namely language learners.

1.1 MT in the classroom

There is a growing literature on the impact of MT in general on the language classroom (see Somers (2001) for an overview), including a series of Workshops at various conferences.[1] Much of the focus is on what trainee translators (or language learners as potential professional translators) should learn about MT, and how MT can be taught to computational linguists. There are also contributions suggesting how MT can be used as a kind of computer-assisted language learning tool. Of interest are approaches which seek to exploit the weaknesses of MT to illustrate the differences between languages, or to heighten learners' appreciation of matters of grammar and style in both languages

[1] See "MT in the classroom" bibliography at http://www.co.umist.ac.uk/~harold/teachMTbibl.html. Dedicated conferences are the *MT Summit VIII Workshop on Teaching Machine Translation*, Santiago de Compostela (2001), *6th EAMT Workshop Teaching Machine Translation*, Manchester (2002), and *MT Summit IX Workshop on Teaching Translation Technologies and Tools (T4) (Third Workshop on Teaching Machine Translation)*, New Orleans (2003).

(Richmond, 1994; Anderson, 1995; McCarthy, 2004; Niño, 2004). However, such uses carry with them the danger that students, particularly beginners, cannot readily identify examples of bad usage, and have a not necessarily justified "trust" in the accuracy of computer output.

The focus of this paper is somewhat different: we are interested in the inappropriate use of FOMT by students seeking a quick way of completing their translation assignments. Indeed, better teaching of what MT is and how it works might be one way to combat this problem, but for the moment we want to explore ways of detecting excessive reliance on FOMT in work submitted by language students. It is shocking to consider that the standard of translation achieved by FOMT might be worthy of a C grade – a moderate pass – at 'A' level.[2] Coupled with the fact that there is a growing move towards coursework-based syllabuses in 'A' levels, it is clear that we need some way of detecting improper use of FOMT by students.

1.2 Use of MT output by language learners

That the availability of FOMT could pose a problem for language teachers is recognized in a thoughtful article by Brian McCarthy (2004). After briefly discussing the role of translation in the language-learning curriculum, suggesting some ways in which the web can be used as a general resource for translators, and suggesting some positive uses of MT in language teaching, particularly as a means of illustrating "translation traps", McCarthy focuses on the "instructional drawbacks" of FOMT.

As he suggests, FOMT "… impacts negatively on the teaching of translation when students simply feed the […] passage they have been given as an assignment through the translation service and submit the […] output for assessment. Motivation for this course of action can vary." Among the causes are "lack of time, lack of energy, or lack of imagination, coupled with a lack of scruples or a lack of linguistic insight".

Submitting output from FOMT for assessment is bad for a number of reasons: it is unfair to students who have invested the intellectual effort and time into producing an original translation; a translation produced with no intellectual input has no instructional value; and it is therefore a waste of the teacher's time to correct it.

McCarthy goes on to report a discussion with his students about strategies for combating FOMT use, and in particular how to penalize it. Interesting though this discussion is, we turn our attention now to the question of how to detect it, which in turn brings us to the question of plagiarism detection.

2. Detecting plagiarism

There is a considerable literature on plagiarism detection, which seems, with the growth of the Internet over the last ten years, to have become a major industry (see Clough (2003) for a good overview). Educators are concerned that students can now too easily complete assignments making inappropriate use of resources found on the Web, whether it be submitting a term paper wholly copied from the Web (perhaps from one of the growing number of "paper mills" and "essay banks"), or more subtle cutting, pasting, combining and editing of several sources without due acknowledgement. There are now numerous services and software packages available which will search the Internet to try to find sources that have been plagiarised, using a number of text similarity measures, to which we will return below. Others will compare sets of documents with each other in order to detect *collusion*, a type of plagiarism where students submit essentially identical assignments because they have worked together on them. This is particularly a problem that is of interest to Computer Science teachers, who were looking at ways of detecting plagiarism in programming assignments long before the Internet came along (e.g. Ottenstein, 1976).

Plagiarism detection has some affinities with and shares some of the techniques of several other branches of computational linguistics and linguistic computing: stylometry and authorship attribution, forensic linguistics, document classi-

[2] This is according to teachers participating in the experiments described below, including one who is an examiner for 'A' levels. These are "advanced level" exams taken by 16-year olds in England and Wales to determine suitability for entry into university. Students usually specialize in three or four subjects at 'A' level.

fication, information retrieval, corpus linguistics.

Our particular interest has two characteristics which make the standard approaches to plagiarism detection less relevant. First, we know beforehand the text (or small group of texts) which we want to check their work against (henceforth, the "source text"). Second, when students do a translation assignment, it is reasonable to expect that there will be textual overlap between their work, corresponding to the range of acceptable translations. So we need to find a way of measuring *excessive* similarity to the source text, and/or perhaps similarity to specific portions of it.

For this reason, we find the related work on *legitimate* reuse of text to be of more relevance, typified by the METER project (Clough et al., 2002), concerned with journalists' use of news agency text.

Plagiarism detection methods are mostly based on string similarity measures ranging from simple vocabulary profiling measures, through string sequence similarity measures to attempts to profile the semantic similarity of texts. In the experiments to be described here, we concentrate on a range of word-counting measures which can be easily implemented and are more or less language-independent.

In the next sections we describe a series of pilot experiments in which we compare 'honest' translations with translations of the same text, derived from FOMT output. We describe a number of measures with which we try to distinguish the two translations.

3. Experiment

Our application is a special case of plagiarism detection, more like Clough et al.'s legitimate reuse detection problem. Considering the nature of student translations, and the (somewhat variable) quality of MT, we can expect the difference between a legitimate but flawed student translation and the inappropriate use of FOMT to be quite subtle. MT systems typically adopt a structure-preserving strategy to translation, but so do students. Furthermore, depending on their level of expertise, students may well make the same sort of lexical choice error (for example, due to too hasty dictionary consultation) as an MT system, and grammatical errors such as incorrect agreement or choice of tense. It is conceivable that the best detection method will somehow home in on the errors that MT makes that no human, however inept at translation, would make.

We could also hope and expect our measures to deliver a scale of values, indicating to the teacher the likelihood that there should be an investigation. It is fairly obvious that the translation that is an exact copy of the FOMT output will be easy to detect; likewise a translation with just one or two words changed, or one or two sentences. Beyond that, the question becomes much more interesting, especially considering also that, arguably, getting students to correct MT output is pedagogically a legitimate and useful exercise (cf. Belam, 2003; Niño, 2004).

This section describes a series of three related experiments with students at various levels and a variety of languages. For our experiments we need some examples of legitimate 'honest' translations, and some examples of lightly post-edited FOMT output. For obvious reasons it would be difficult to get genuine examples of the latter, so we devised a means of generating parallel sets of translations done with and without the 'help' of FOMT. The students were asked to perform two tasks with the text. One task was to translate it into English using 'normal' resources (dictionaries, grammar reference books); the other was to take the Babelfish translation and 'tidy it up' as much as possible in a strictly limited timeframe. Henceforth we refer to these as 'honest' and 'derived' translations respectively. Ideally, we would have liked to get both types of translations from all the students, and also split the students into two groups, so as to control for order of task completion: this was only possible with one group, where we had to compromise on the translation task for lack of time. For the other groups we have a large number of derived translations, and a rather smaller number of honest translations done by the same students.

The first group consisted of ten students studying Italian at the University of Manchester. Because of the shortage of time available, for the honest translation task the text was split into two halves, and students worked on one or other portion. All students did the derived translation. The second group of students was made up of

ten intermediate Spanish students enrolled at the Centre of Continuing Education at Manchester University. Half of them did the honest translation, the other half the derived translation. The third group eventually consisted of five sixth-form students (years 11 and 12) studying German. As part of a language-lab class, a much larger group of about 45 students studying French, German, Spanish and Russian spent about 40 minutes doing derived translations, of which about half completed a significant amount of work. The students were encouraged to provide an honest translation several weeks later (with the Christmas break in between), but unfortuantely only five students did so. The students had studied German for several years, but the English foreign-language syllabus does not include traditional translation until 'A' level, so for them translation was a relatively new task.

In each experiment we used the same source text, a short English text (224 words in 14 sentences) from a website[3] and used the AltaVista Babelfish service to translate it into the various languages. In what follows we will make a distinction between 'derived' translations resulting from post-editing the Babelfish output, and 'honest' translations which are done in the traditional manner.

Some examples of the students' work are shown in (1): (1a) shows the original English, (1b) the Babelfish translation into Spanish, (1c-d) two derived translations, and (1e-f) two honest translations.

(1) a. I want to run at the 2012 Olympics for South Africa.
b. *deseo funcionar en las 2012 Olimpiadas para Suráfrica.*
c. *deseo correré en las 2012 olimpiadas para Suráfrica..*
d. *Deseo correr en las Olimpiadas de 2012 para Sudáfrica.*
e. *Yo quiero correr en los juegos de 2012 por sud Africa.*
f. *Quiero correr en las olimpiadas 2012 por Africa del sur.*

[3] "Stars of Singapore visit east London school", www.london2012.org/en/news/archive/2005/October/2005-10-05-13-42.htm

4. Measures and results

4.1 Simple word counts

Early attempts at computational stylometry focused on simple statistics based on word frequency counts (cf. Baayen, 2001). Measures such as type–token ratio, and other measures of vocabulary richness are not appropriate for our task, as the texts are too short. However, from this field comes the idea of counting *hapax legomena* (HL; lit. 'once said', i.e. words occurring once in the text, also termed "singletons"): the idea is that a significant overlap in use of infrequent words might suggest copying. This is the basis of the *CopyCatch* program (Woolls and Coulthard, 1998), in which it is claimed that an overlap of 70% is suspicious. We measure HL overlap by counting the percentage of singletons in the source text which are also singletons in the target. Words occurring exactly twice – *dis legomena* (DL; "doubletons") – also have a distinctive distribution, so we measure DL overlap too. By extension, we propose a measure taking into account all "*n*-letons": if we consider the different totals for all the frequencies, we can calculate a Euclidian distance measure *F* as in (2),

(2) $$F(s,t) = \sqrt{\sum_{i=1}^{n} \left(f_i^{\,s} - f_i^{\,t}\right)^2}$$

where f_i^x is the number of words occurring with frequency i in text x, n being the frequency of the most frequent word.

Looking at the frequencies of individual types in the two texts, two further measures of text similarity suggest themselves. The first is the percentage of words that have exactly the same frequency (SF) in the two texts. The second is again a Euclidian distance *E*, this time based on the total frequencies of the words in each text, as in (3),

(3) $$E(s,t) = \sqrt{\sum_{w \in (s \cup t)} \left(f_w^s - f_w^t\right)^2}$$

where f_w^x is the frequency of occurrence of the word w in text x.

Figure 1 shows the scores for these five measures in graphic form. Black symbols show derived translations, white symbols honest translations. The shapes indicate the languages: triangles for Spanish, circles for Italian, squares for German. The figure shows that of these five

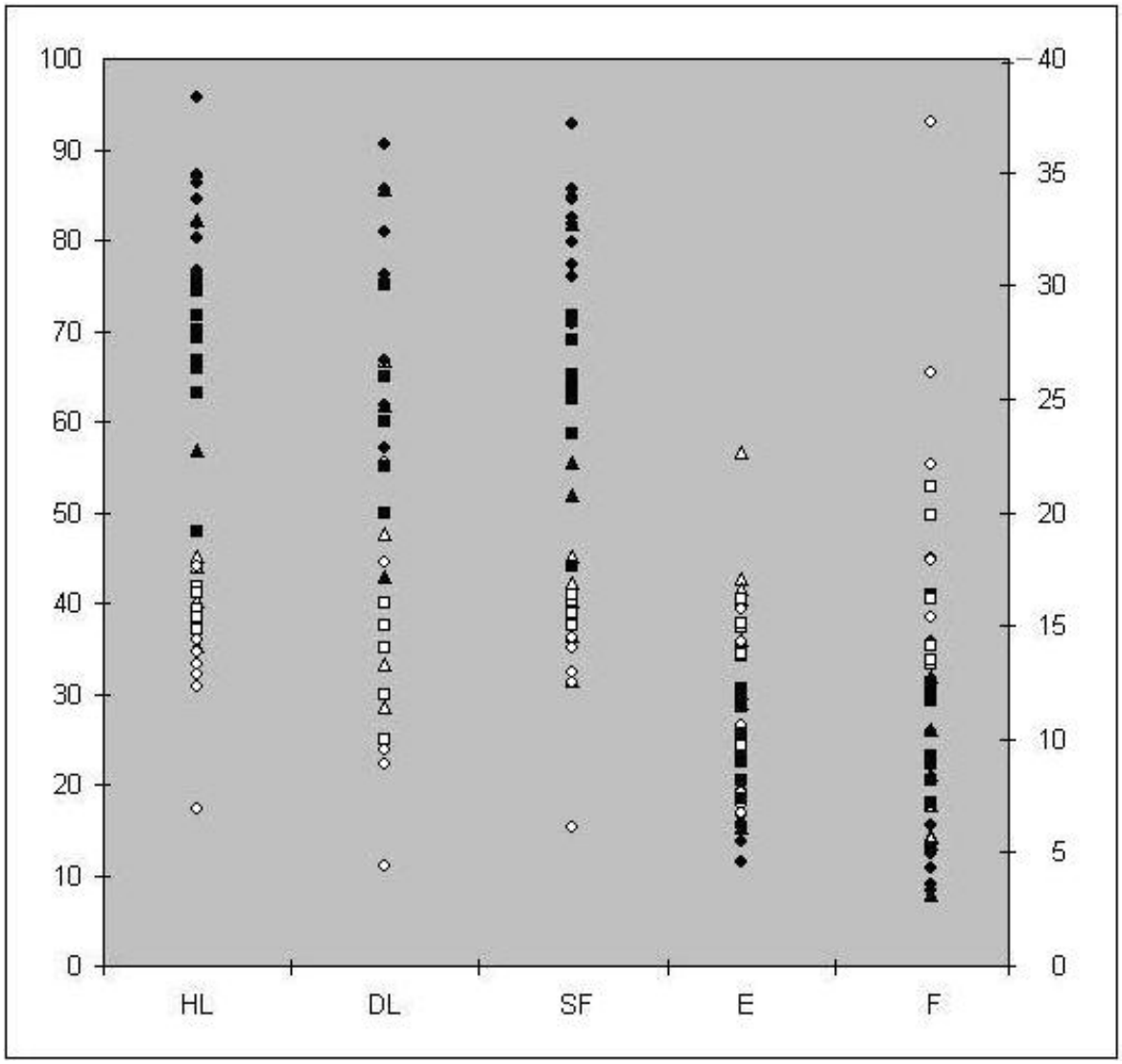

Figure 1. Word-count measures for comparison of derived and honest translations with Babelfish text. HL, DL and SF are expressed as percentages. *E* and *F* are plotted on the right-hand axis. Black symbols show derived translations, white symbols honest translations. The shapes indicate the languages: triangles for Spanish, circles for Italian, squares for German.

proposed measures, HL works best: high scores indicate derived translations, low scores suggest honest translations. Apart from two texts, there is a clear gap of nearly 20 points between the derived and honest translations. With a *t* score of 2.533, the result is statistically significant at $p < 0.02$. We could conclude that any HL score above 50% was suspicious. The DL score is similar in principle, but not so decisive: there is quite an overlap in the range of scores, and one of the honest translations scores as highly as 66%. At $t = 1.279$ the result is not statistically significant at any reasonable level. The SF score based on word frequency again reflects a tendency, but with overlap around the middle of the score range, and the *t* score of 2.302 is significant at $p < 0.05$. The Euclidian distance measures *E* and *F* are not able to distinguish at all, and can reasonably be abandoned. Although *F*, a third measure based on the frequency distribution, shows some clustering (this time a low score indicates similarity, and thus suspicion), there is considerable overlap. The *t* scores of 0.819 for *E* and 0.836 for *F* are not statistically significant.

4.2 Comparing word sequences

An intuitive way of detecting plagiarism is to look for common sequences of words, and indeed this has been the basis of several approaches. Searching for overlapping *n*-grams has been used for example by Brin et al. (1995), Heintze (1996), Shivakumar and Garcia-Molina

(1996) and Lyon et al. (2001). A use of *n*-gram matching that is very familiar in the world of MT evaluation is in the BLEU (Papineni et al., 2002) along with Doddington's (2002) derived NIST algorithm. Both these measures essentially give a weighted precision score based on the number of *n*-grams common to both source and target texts. In the `mteval` implement-ation[4] *n*-grams up to $n = 9$ are included. While the idea of *n*-gram matching against an oracle translation is somewhat controversial for MT evaluation, it seems to offer a good platform for evaluating the similarity of two texts.

A simpler measure of text similarity based on word sequences is of course Levenshtein (or string-edit) distance (LD) (Levenshtein, 1965). Implementations can vary as to whether they count only substitutions, insertions and deletions ("indels"), or also count transpositions, mergers and expansions. Also, segments can be treated as strings of words or strings of characters. For our application, we calculate the LD in its simplest form (substitutions and indels) for each segment taken as a string of words, and provide an average over the individual segment scores.

Both the BLEU/NIST algorithms and the LD rely on the two texts being sentence-aligned. Fortunately, student translations typically follow the structure of the source text fairly closely, and students generally translate sentence by sentence. MT systems certainly do.

Figure 2 shows the LD, BLEU and NIST scores for our data. Again, black symbols show derived translations, white symbols honest translations, while the shapes indicate the languages.

All three measures show a clear separation of the honest texts from the derived translations, though with the NIST scores the top-scoring honest translation is very close to the bottom-scoring derived translation. The results for LD and BLEU are statistically significant at $p < 0.02$, while NIST is slightly weaker with significance at $p < 0.05$. LD has the advantage that it can show us how much and exactly where the texts differ. Figure 3 shows the individual LD scores (for the Italian data) on a sentence-by-sentence basis, expressed as a percentage of the number of words in each sentence in the original text. The actual length of the sentence in the Babelfish translation is important in interpreting the importance of a high or low score. For example, even the honest translations have an LD of 0 for the final 3-word sentence (the date), and both honest and derived translations show a high percentage change for the first sentence, the 7-word title.

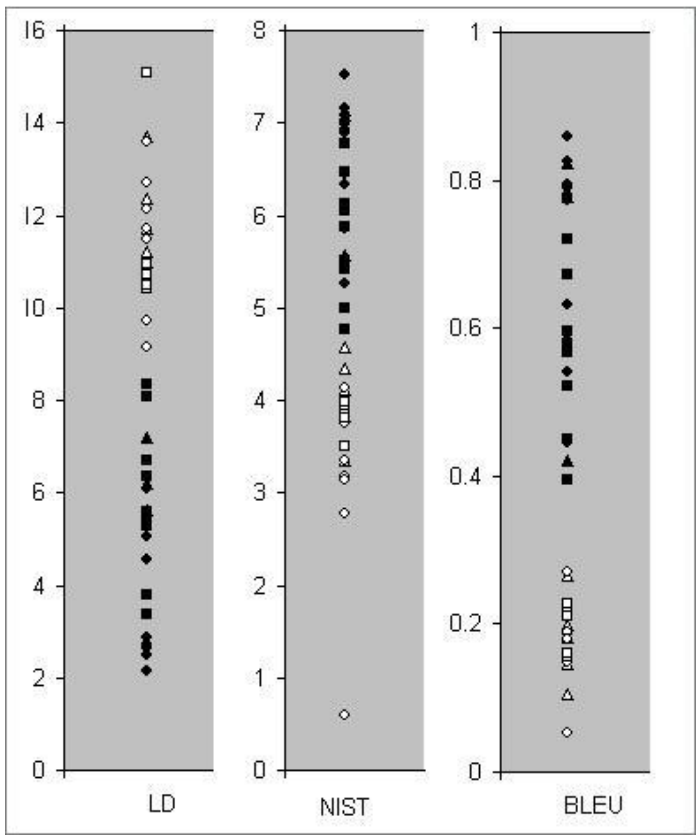

Figure 2. LD, BLEU and NIST scores for comparison of derived and honest translations with Babelfish text.

5. Conclusions

Although these results are based on a small experiment with a few students, they do suggest that there are a number of measures that can indicate that a translation is suspiciously similar to a free online version, namely: HL (relative distribution of singleton lexical items), SF (percentage of words having the same frequency), LD (a measure of the difference between the texts, sentence by sentence), and the

[4] Downloadable from www.nist.gov/speech/tests/mt/resources/scoring.htm.

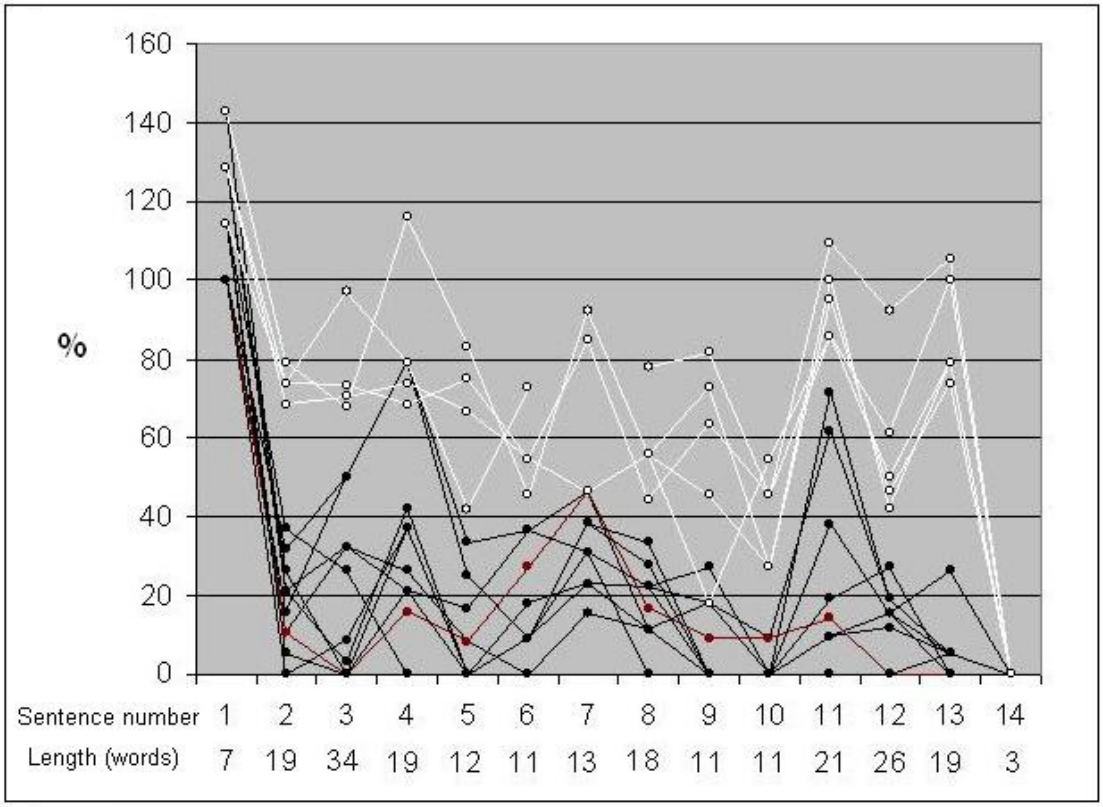

Figure 3. Sentence-by-sentence Italian LD scores expressed as a percentage of sentence length.

BLEU and NIST scores (measures of *n*-gram overlap). For our purposes this is sufficient, as it will signal to the teacher that the work *might* be plagiarized, and should be looked at more closely. What the teacher does with the student caught plagiarizing is of course another issue, as is the problem of students using FOMT to produce texts which they were meant to compose in the foreign language. It seems to us plausible that the mistakes made by MT systems are sufficiently different from those made by language learners to permit some sort of automatic detection, but this would depend on techniques of computational stylometry rather than plagiarism detection – a topic for a further study perhaps.

Acknowledgments

Grateful thanks to Mr Ian Leverton, his staff and students at The Manchester Grammar School. Our gratitude also to Italian students in the LEAP programme and Spanish students from the Centre for Continuing Education, both in the University of Manchester. And thanks are due to Tim Morris for help with the statistics.

References

ANDERSON, Don D. (1995) Machine translation as a tool in second language learning. *CALICO Journal* **13**.1, 68–97.

BAAYEN, R. Harald (2001) *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

BELAM, Judith (2003) "Buying up to falling down": A deductive approach to teaching post-editing. *MT Summit IX Workshop on Teaching Translation Technologies and Tools (T^4) (Third Workshop on Teaching Machine Translation)*, New Orleans, LA, pages 1–10.

BENNETT, Winfield Scott (1996) Learning from users. *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Quebec, pages 229–231.

BRIN, Sergey, James DAVIS and Hector GARCIA-MOLINA (1995) Copy detection mechanisms for digital documents. *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, San Jose, California, pages 398–409.

CLOUGH, Paul (2003) *Old and new challenges in automatic plagiarism detection*. JISC National Plagiarism Advisory Service, Newcastle-upon-Tyne, Available online at http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf.

CLOUGH, Paul, Robert GAIZAUSKAS, Scott S.L. PIAO and Yorick WILKS (2002) METER: MEeasuring TExt Reuse. *ACL-02: 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pages 152–159.

DODDINGTON, George (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *HLT 2002 Human Language Technology Conference*, San Diego, CA.

FLANAGAN, Mary (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, Quebec, pages 192–197.

HEINTZE, Nevin (1996) Scalable document fingerprinting. *Proceedings of the Second USENIX Workshop on Electronic Commerce*, Oakland, California.

LEVENSHTEIN, Vladimir Iosifovich [= ЛЕВЕНШТЕЙН, Владимир Иосифович] (1965) Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академий Наук СССР* **163**.4:845–848. Appeared as: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **10** (1966):707–710.

LYON, Caroline, James MALCOLM and Bob DICKERSON (2001) Detecting short passages of similar text in large document collections. *2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA, pages 118–125.

MCCARTHY, Brian (2004) Does online machine translation spell the end of take-home translation assignments? *CALL-EJ Online* **6**.1. Available at www.clec.ritsumei.ac.jp/english/callejonline/9-1/mccarthy.html.

MIYAZAWA, Shinichiro, Shoichi YOKOYAMA, Masaki MATSUDAIRA, Akira KUMANO, Shuji KODAMA, Hideki KASHIOKA, Yoshiko SHIROKIZAWA and Yasuo NAKAJIMA (1999) Study on evaluation of WWW MT systems. *Machine Translation Summit VII '99: MT in the Great Translation Era*, Singapore, pages 290–298.

NIÑO, Ana (2004) Recycling MT: A course on foreign language writing via MT post-editing. *7th Annual CLUK Research Colloquium*, Birmingham, pages 179–187.

OTTENSTEIN, Karl J. (1976) An algorithmic approach to the detection and prevention of plagiarism. *ACM SIGCSE Bulletin* **8**.4, 30–41.

PAPINENI, Kishore, Salim ROUKOS, Todd WARD and Wei-Jing ZHU (2002) BLEU: A method for automatic evaluation of machine translation. *ACL-02: 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pages 311–318.

RICHMOND, Ian M. (1994) Doing it backwards: Using translation software to teach target-language grammaticality. *Computer Assisted Language Learning* **7**, 65–78.

SHIVAKUMAR, Narayanan and Hector GARCIA-MOLINA (1996) Building a scalable and accurate copy detection mechanism. *DL'96: First ACM Conference on Digital Libraries*, Bethesda, MD.

SOMERS, Harold (2001) Three perspectives on MT in the classroom. *MT Summit VIII Workshop on Teaching Machine Translation*, Santiago de Compostela, pages 25–29.

WOOLLS, David and Malcolm COULTHARD (1998) Tools for the trade. *Forensic Linguistics* **5**, 33-57.

YANG, Jin and Elke D. LANGE (1998) SYSTRAN on AltaVista: A user study on real-time machine translation on the Internet. In David Farwell, Laurie Gerber and Eduard 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, pages 275–285.

YANG, Jin and Elke LANGE (2003) Going live on the Internet. In Harold Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam, Benjamins, pages 191–210.