

Exploiting the WWW for MT teaching

Lorna Balkan

Dept. of Language and Linguistics,
University of Essex,
Wivenhoe Park, Colchester, Essex, UK.
balka@essex.ac.uk

Abstract

This paper gives an overview of what resources, including software tools, reference material and course material, are currently available on the web for teaching machine translation, and discusses where to find these resources. It makes some suggestions as to how these resources and access to them can be enhanced in the future.

1 Introduction

This paper gives an overview of what the World Wide Web (henceforth simply “the web”) currently offers in the way of resources for teachers and learners of machine translation (MT), and discusses how its potential can be further exploited in future. Given the problems of discovering resources on the web, the study does not claim to be exhaustive. In particular, it is confined in the main to what resources are available in English. However, hopefully the short overview that follows will give some flavour of the type of material that is out there and where to find it. The difficulty involved in finding these resources, and the scarcity of resources in some areas, are then used to suggest some possible ways forward for MT teaching on the web, and steps the research community can undertake to promote them.

2 The Web as a Teaching Resource

The value of the web as an educational resource is being increasingly recognised. In Britain, there are some excellent web resources particularly for school children, and web-based learning (or “e-learning” as it is sometimes referred to) is gradually finding its way into the classroom. Different ways of exploiting the web have been explored. The web can be used first and foremost as an information resource, as a sort of virtual library, where useful books, articles, reviews, etc, relating to the subject being taught can be found and stored. More active uses of the web include online courses, with the possibility of providing automatic feedback to students.

Web-based learning has a number of important advantages over traditional methods of accessing information, providing as it does fast, usually free, access to reference material which might otherwise be difficult and/or expensive to track down. This is the more attractive, given the financial constraints under which university libraries, for example, have to operate. Online access also makes materials

available to a wider body of students, since distance is no barrier.

Those who promote e-learning are, however, the first to admit its drawbacks as well as its disadvantages. Certainly, the web is useful for tracking down useful resources, but given the millions of web pages, just how do you find the resources that you need? Apart from using traditional search engines (eg. Altavista) , a number of education or subject portals are available that attempt to do the hard work for you. In Britain for example, there is the National Grid for Learning (<http://www.becta.org.uk/start/ngfl.html>), an education portal, and the more specific Languages, Linguistics and Area Studies subject centre (<http://www.lang.ltsn.ac.uk/>) which provides information, resources and links for teachers of linguistics and related subjects. These portals have the added advantage of screening web sites so that the user can be sure of finding reliable information, since anyone who has used the web will know that it contains as much bad information as good. Another issue with web-based learning concerns the writing of online course material. It is not just a question of transferring traditional lecture notes to web pages, but a skill in itself. The development costs can be very high, and are usually only justified where there are a large number of students.

3 MT for whom?

MT as a subject is still not widely taught, but there is reason to believe that demand for it may be increasing. This is suggested by two factors. Firstly, the ever increasing availability of free online MT resources (see Section 5 below) is making MT available to more and more people, some of whom will certainly be interested in learning more about it. Secondly, there seems to be an ever increasing range of multilingual applications (e.g. multilingual information retrieval systems) which incorporate MT as a component part, and researchers and developers in these areas are also interested in finding out about MT technology.

Students of MT are therefore set to increase in diversity, as well as in number. A survey conducted in 1996 on tools and techniques for MT teaching (Balkan et al., 1997) revealed that over 50% of students who studied MT were computational linguists. Whether this reflected the true state of affairs at that time is hard to say, since the questionnaire on which the study was based received few replies from translating/interpreter training schools or departments. A quick look at course descriptions for translating courses in Britain in 2001 reveals that the teaching of MT or Machine Aided Translation (MAT) features in a good number of them. This is surely a reflection of the better acceptance of MT/MAT in the translator community.

It is clear that students with different backgrounds and interests have different needs with respect to MT teaching. A student of computational linguistics is likely to approach the subject from a completely different angle from, say, a translation student. While the former is likely to be well-versed in both computers and linguistics, and therefore able to handle complex descriptions of systems and translation engines, the latter will be interested in MT mainly from the point of view of the user, and no knowledge of either computers or linguistics can necessarily be assumed. The different requirements of these different types of user should therefore be borne in mind when discussing web-based MT teaching.

4 Locating Resources

Where does one start in looking for web-based resources for MT teaching? In the absence of a portal dedicated to MT, finding resources and materials for MT teaching is often a matter of searching the web using a general search engine, and following links on pages that are found in this manner. However, this is a time-consuming and not always efficient method. Typing “machine translation” on the Altavista website (<http://www.altavista.com/>) produced 31,195 hits (on 9/4/01)!

A more informed search might start with a dedicated MT web-site, and search from there. A good place to start is the website of the European Association for Machine Translation (EAMT) (<http://www.eamt.org/>) or one of its sister associations. The EAMT website manages an email list, which MT students may be encouraged to join to receive up-to-date information about forthcoming conferences on MT, etc. The site also contains reports of interest to students of MT (see section 5 below), and provides further links to other associations, organisations and research centres that specialise in MT and/or NLP. Important amongst these are the Association for Computational Linguistics (ACL) (<http://perun.si.umich.edu/~radev/u/db/acl/html/>) and the European Network for Language and Speech (ELSNET) (<http://www.elsnet.org/>). The ACL website contains,

amongst other things, a directory of subject-specific resources, including some eight entries for Machine Translation. These consist in the main of pointers to other MT related sites. Exploration of the ACL website is made easier by a search facility that returned 147 hits for the query “machine translation” (on 9/4/01). The ELSNET website is well organised and contains much of interest to the MT student, including an online readable version of its newsletter, which contains up-to-date information about NLP developments, including MT, a directory of NLP and speech tools, and a directory of books and publications. It also has an as-yet-incomplete directory for special topics, including MT, which has only two entries to date.

Another possible starting place is Human Language Technology (HLTCentral) (<http://www.hltcentral.org>) gateway to speech and language opportunities on the web. It covers news, R&D, technological and business developments in the areas of speech, language, including MT and produces an online publication called *le Journal*. The site also contains useful reports on language technology as a whole.

At a local level, the Natural Language Translation group of the British Computer Society (<http://www.bcs.org.uk/siggroup/sg37.htm>) provides useful links. They have a list of books and journals devoted to MT and a list of language and linguistics related email lists.

Other possible starting places are sites devoted to language in general, not just language technology. The Human Language Page <http://www.june29.com/HLP/>, for example, contains much that may be of interest to an MT student, including a list of dictionaries and a directory of linguistic resources, mainly links to other centres. An important site dedicated to languages in general is Linguist List (<http://www.linguistlist.org/>) which contains links to a large number of linguistic resources, including project and research sites and collections of papers. Linguist List also manages an email list, the archives of which can be searched.

Other sites that contain extensive indexes of linguistic resources include SIL International (<http://www.sil.org/linguistics/topical.html>), the University of Rochester (<http://www.ling.rochester.edu/linglinks.html>), Stanford University (<http://www-nlp.stanford.edu/links/linguistics.html>) and Blackwells the publisher (<http://www.blackwellpublishers.co.uk/LINGUIST/default.htm>).

5 Tools for MT Teaching

The most important tools as far as MT teaching is concerned are the MT/MAT tools themselves. By MAT we mean any linguistic resource that aids the translation process, while

leaving the onus of the translation task to the user. (We contrast this with our definition of MT, where it is the machine that does the bulk of translation work, producing a raw translation that the user then post-edits.) MAT tools include dictionaries, terminologies, thesauri, bilingual concordances, and translation memory systems.

There are so many online dictionaries that it is not possible to access them all from a central site. However, some sites contain useful compendia, such as <http://www.yourdictionary.com/> (Dictionaries and thesauri), and http://www.fbi.fhkoeln.de/labor/bir/thesauri_new/theslang.htm for thesauri. Another useful MAT tool is the bilingual concordance program of the Canadian Hansard (1986-1993) <http://www-rali.iro.umontreal.ca/TransSearch/TS-simple-uen.cgi>, which has an interface in both French and English. We are not aware of a free online translation memory, but the idea of how they work can be gleaned from descriptions and demos of such systems, e.g. the Trados web site at <http://www.trados.com/>.

These MAT tools can give the student some idea of the problems involved in accessing and using multilingual tools online. The student can explore, for example, how the system deals with accents or capital letters, and how it copes with non-exact matches. The student can be set a text to translate and see how the various tools can be exploited and/or compared.

As to MT proper, there seems to be an ever-increasing number of commercial MT systems that are becoming available online. Several sites provide a compendium of online MT systems, including <http://www.word2word.com/free.html>. A problem with these sites is their lack of completeness and scarcity of information about the systems. The latter factor limits their usefulness as teaching aids. The EAMT website offers some help in these areas, in that they produce a compendium of MT systems that is updated every four months, and that is downloadable from their site at no charge to members and for a small cost to non-members of EAMT and sister organisations. Also available from the EAMT website <http://www.eamt.org/resources/index.html> is a fairly recent annotated list of free online commercial MT products, produced by Laurie Gerber.

Research MT systems are harder to locate, but the MT system KIT-FAST is available for downloading from <http://wave.cs.tu-berlin.de/~ww/mtsystem.html>. Like commercial systems, research systems need to be supported by documentation to be of use for teaching purposes. They also have to be easy to install.

There are a number of other tools available online that may also be of interest to the student of MT to the extent that they may be constituent parts of MAT or MT products. This category includes morphological analysers, parsers, taggers,

etc. A good place to look for such tools are repositories such as the Natural Language Software Registry <http://www.dfki.de/lt/registry/>, the European Language Resources Association (ELRA) <http://www.icp.inpg.fr/ELRA/> which distributes speech, text and terminology resources and tools, and the Linguistic Data Consortium <http://www ldc.upenn.edu/>, which collects and distributes corpora, speech databases, lexicons, etc.

Carnegie Mellon University has an MT tools repository (<http://www.cs.cmu.edu/afs/cs/project/airepository/ai/areas/nlp/mt/0.html>) which contains just two items, COGNATE, which identifies words in a bilingual text, and MTRAN, an experimental MT system, both of which can be downloaded.

Other useful sites for locating NLP software include Penn University (<http://www.cis.upenn.edu/~adwait/penntools.html>), Tokushima University http://n106.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html, and Mary Taffet's web page http://web.syr.edu/~mdtaffet/nlp_sites.html#Res_Tools which includes a list of online demos.

For students interested in the evaluation of MT, test suites in English, French and German which were produced as part of the EU-funded TSNLP project are available online from <http://tsnlp.dfki.uni-sb.de/tsnlp/tsdb/tsdb.cgi>.

6 Reference materials

We now consider what reference materials, including books, reports/articles of a general nature, and scholarly articles and papers, are available online.

Apart from sites mentioned in Section 4 above, sites where collections of scholarly papers on MT can be found include the Computing Research Repository (CoRR) (<http://www.acm.org/pubs/corr/>), formerly the Computation and Language E-Print Archive, a searchable archive and distribution server for papers on computational linguistics, natural language processing and related fields dating from 1993. Searching for "machine translation" in the "abstract" field of the computer science archive for all years produced 67 results (on 9/4/01).

Also useful are sites dedicated to particular MT projects, including the following projects: Verbmobil <http://verbmobil.dfki.de/cgi-bin/verbmobil/htbin/doc-access.cgi>, the SRI Spoken Language Translator <http://www.cam.sri.com/tr/slt.html>, KANT <http://www.lti.cs.cmu.edu/Research/Kant/>, Pangloss <http://www.lti.cs.cmu.edu/Research/Pangloss/Home.html>, Mikrokosmos <http://crel.nmsu.edu/Research/projects/mikro/index.html>, and CAT2 <http://www.iai.uni-sb.de/global/cat-docs.html>. The EUROTRA reference manual is available online from <http://www.iai.uni-sb.de/REFMAN/pap-refm.html>.

Two sites of direct relevance to the evaluation of MT deserve a mention. The ALPAC report (ALPAC, 1966) is available from

<http://www.nap.edu/books/ARC000005/html/>, and the project International Standards for Language Engineering (ISLE), which is developing a theory about the methodology for evaluating Natural Language Processing/Computational Linguistics applications in general, has produced a specific framework for classifying evaluations of MT in particular, which can be viewed at <http://www.isi.edu/natural-language/mteval/>.

In Balkan et al. (1997) two books/reports available online were mentioned that were of particular interest to the MT teaching community. They were Arnold et al. (1994), available from <http://clwww.essex.ac.uk/MTbook/> and Cole et al. (1996) available from

<http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>. Both these publications continue to be available online. Arnold et al. (1994) is an introductory book and assumes no prior knowledge of either linguistics or computing. Cole et al. (1996) has two useful sections on CAT and MT, which are accessible to non-specialists. Another report that is of interest to students of MT has since become available online – Hovy et al. (1999), which can be viewed at <http://www.cs.cmu.edu/~ref/mlim>. Chapter 4, edited by Bente Maegaard looks at machine translation. Other chapters, including Chapter 1 on multilingual resources, and Chapter 8 on evaluation and assessment techniques, are also of potential interest.

There are a number of very useful general articles on MT available online. John Hutchins' web site (<http://ourworld.compuserve.com/homepages/WJHutchins/>) contains accessible papers on the history, etc. of MT, while Language Partners International (<http://www.languagepartners.com/reference-center/index.htm>) has a number of useful articles aimed at MT users, including "An Introduction to Computer Aided Translation (CAT)" and "How To Select the Right CAT Tool Solution".

7 Other Teaching Materials

In this category we include courses, syllabi, and other educational resources. This is the most difficult area to review, since there is no obvious place to search, except in the web sites of particular institutions and individuals. However, this situation seems set to change. JEWELS <http://www.hltcentral.org/page-799.0.shtml> is an as-yet-incomplete EU funded website for educational materials in Language and Speech. Its aim is to provide recommendations on higher education curricula in Language and Speech and information about courses, as well as a searchable database for tools that can be used in education.

Of interest here is not just to know what online courses are available, but to have guidelines for creating online materials of one's own. Although there are many online courses already available for the teaching of NLP and related topics (see for example <http://instruct.uwo.ca/gplis/677/thesaur/main00.htm#contents> for a tutorial on thesaurus construction) it is more difficult to find material of direct relevance to MT. Teaching material from a workshop on statistical MT, including Kevin Knight's Statistical MT tutorial workbook, is however available from <http://www.clsp.jhu.edu/ws99/projects/mt/index.html>.

Guidelines for producing web-based materials can be found, amongst other places, in Lee et al. (1999), available from <http://info.ox.ac.uk/jtap/reports/teaching/index.html>.

8 The Way Forward

In summary, there is much of interest out there for the MT teacher and student, if only one knows where to find it. As resources proliferate, we need a systematic way of finding them. What would be really useful would be a central gateway for MT, which would keep an up-to-date record of tools and resources, and other information on MT.

The need for keeping track of and sharing resources is being increasingly recognised in other areas of linguistics. Initiatives such as the forthcoming ACL/EACL 2001 Workshop on Sharing Tools and resources for Research and Education are indicative of this recognition. Already much has been done, such as the setting up of the JEWEL project (see Section 7 above), and ELRA (see Section 5). Efforts such as these need to be continued and supported.

Ideally an MT portal would contain not just a list of resources, but one that has been properly annotated and preferably indexed to allow for easy browsing and searching. Retrieval could be on the basis of language, topic (e.g. statistical MT), type (e.g. software tool, book, etc.), and level (e.g. technical, non-technical). As discussed in Section 3 above, not every student of MT will be able to tackle material of a very technical nature.

Finally, this overview of MT teaching resources has revealed areas where more resources or information are required. Lack of information about commercial MT systems, highlighted as a problem in Balkan et al. (1997) continues to be a problem. It would also be helpful if more research systems, or demos at least, were made available.

While new research papers are increasingly finding their way online, this is done on a rather unsystematic basis. It would also be useful if some of the more "classic" papers on MT, such as the ALPAC report (ALPAC (1966)) could be made available online.

While there are many issues to be resolved, such as copyright issues and format and metadata standards, it is to be hoped that initiatives such as the current workshop can be instrumental in bringing about some of the above recommendations.

References

- ALPAC (1966). Language and Machine: Computers in Translation and Linguistics. Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Research Council), Washington, D.C, National Academy of Sciences.
- Arnold, D., Balkan, L., Humphreys, R.L., Meijer, S., and Sadler, L. (1994). Machine Translation: an Introductory Guide. Manchester: NEC Blackwell.
- Balkan, L, Arnold, D. and Sadler, L. (1997). Tools And Techniques for Machine Translation: a Survey. Available from <http://clwww.essex.ac.uk/group/projects/MTforTeaching/>.
- Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., and Zue, V. Survey of the State of the Art in Human Language Technology (1996). A study sponsored by the US National Science Foundation, Directorate XIII-E of the Commission of the European Communities, and the Center for Spoken Language Understanding, Oregon Graduate Institute.
- Hovy, E., Ide, N., Frederking, R., Mariani, J., and Zampolli, A. (1999). Multilingual Information Management: Current Levels and Future Abilities. A report commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defense Advanced Research Projects Agency.
- Lee, S., Armitage, S., Groves, P., and Stephens, C. (1999). Online Teaching: Tools & Projects. A report commissioned by the Joint Information Systems Committee (JISC) Technology Applications Programme, UK.