Teaching and Assessing Empirical Approaches to Machine Translation

Andy Way & Nano Gough

School of Computing,
Dublin City University,
Dublin, Ireland.

Email: {away,ngough}@computing.dcu.ie

Why Teach Empirical Approaches to MT?

- These techniques are now *mainstream* in the research field;
- Important that the tools and techniques in these paradigms be taught to potential future researchers and developers in University courses.

Status Quo: they're not (really) being taught

- Compare the situation with Statistical NLP, where many dedicated courses can be found.
- In MT, there are few, if any courses on Empirical Approaches.

We couldn't find *any* on the Web. What *is* there are (brief) mentions of the techniques in a couple of classes in modules dealing with wider issues in MT:

- MSc. in MT, CALL and NLP at UMIST, UK;
- MSc./Ph.D. Program in Language and Information Technologies at CMU, Pittsburgh.

and in Empirical NLP:

- Programmes in Computer Science, ISI, CA;
- Undergraduate Study in Computer Science at Brown University, Providence, RI;
- Postgraduate programmes in Computer Science at UMIACS, MD.

Similar situation w.r.t. textbooks. Dedicated volumes include:

- Statistical NLP:
  - Charniak, 1993;
  - Manning & Schütze, 1999.
- EBMT:
  - Carl & Way, 2003.
- SMT:
  - ??? (*Statistical MT Workbook*).

Some newer textbooks on NLP/MT address these issues, including:

- Trujillo (1999, Chapter 8) goes into some detail on EBMT and SMT;
- Bowker (2002, Chapter 5) and Somers (2003, Chapter 3) discuss the related area of TM systems;
- Melamed (2001) is geared specifically towards the exploitation of bitexts using empirical methods;
- to a lesser extent, Jurafsky & Martin (2001, Chapter 21) provide some detail on how empirical techniques can be used in MT, as does Somers (2003, Chapters 3 and 8).

Until (Carl & Way, 2003), if you wanted to teach EBMT you had to rely on original papers and survey articles (e.g. Somers, 1999).

If you want to teach SMT, you *still* have to do this (e.g. Brown *et al.*, 1990, 1992; Yamada & Knight, 2001; Soricut *et al.*, 2002).

What's this talk about?

A course taught by us on Empirical Methods in MT. Not intended to be prescriptive:

- *May* serve as a model considering teaching such a module;
- Reports on what worked and—more importantly, what didn't—so others may benefit from lessons learnt.

Courses must be tailored to students' skills and demands

cf. distinction between *users* versus *developers* (e.g. Kenny & Way 2001):

- language students/translators need to know:
  - how to use translation tools (i.e. TM) in their careers as translators;
  - (superficially) about EBMT, and especially the differences between TM and EBMT.
- students of NLP with a specialisation in MT might (realistically) be employed as designers and implementors of such tools in a programming or localisation environment.

Note also that while both EBMT and TM require aligned corpora, for instance, the TS students are far more likely to *use* built-in alignment tools such as Trados *WinAlign*, NLP students may be expected to *develop* their own alignment software.

What's the profile of our students?

Final year undergraduates taking a degree in NLP, who (should) have:

- a strong background in programming;
- good competence levels in formal linguistics and NLP;
- good L2 skills;
- already done another module in RBMT.

So our course is very practically oriented.

By the end of the course, we expect(ed) students to be able to develop a toy EBMT system.

Note again then that our course is geared specifically to a group of final year undergraduate students taking a degree in NLP. It would, therefore, be an inappropriate model for groups of students with differing backgrounds.

Empirical Approaches to MT: Course Content

Taught 3 hours a week lectures and a 2 hour practical session over a period of 8 weeks.

Material delivered in class during one week is put into practice in the labs during the following week.

Course is split into three chunks:

- Alignment (both sentential and word-level);
- EBMT;
- SMT.

Programming language used: Perl.

Introductory part of the course

Students are made aware that statistical language and translation models should be developed from *large, good quality, representative* monolingual and bilingual corpora. We concoct toy corpora which do not fulfill these criteria, and ask students to calculate a number of unigram and bigram probabilities based on data contained in these corpora. It is easy to show that a number of undesirable effects follow when small, unrepresentative corpora are used.

Empirical methods are by and large utilisable for any pair of languages. But sententially aligned bilingual corpora exist only for a few language pairs. To try to overcome this problem, some consideration is given to using the Web as a corpus from which usable bitexts might be extracted (cf. Grefenstette, 1999; Resnik & Smith, 2003).

Sentential Alignment Algorithms

We present some of the major algorithms for aligning bilingual corpora:

- Brown *et al.*, 1991;
- Gale & Church, 1993;
- Kay & Röscheisen, 1993.

In the practical sessions, we get the students to write simple routines to perform sentential-level alignment using:

- relative sentence position;
- relative length of sentence;
- combinations of the two.

These are then gradually improved by adding in *anchors* such as:

- cognates;
- paragraph markers;
- punctuation;
- HTML tags.

Sub-Sentential Alignment

Students are shown how to estimate co-occurrence using Mutual Information. W.r.t. EBMT in particular, other methods of segmentation are also presented, especially *Marker-Based* segmentation (e.g. Veale & Way, 1997; Way & Gough, 2003).

Students are also made aware of the need for extracted sub-sentential alignments to be made more *general*, in order to improve *coverage* and *robustness*. Some of the techniques presented include:

- Extracting transfer rules from examples (e.g. Furuse & Iida, 1992);
- Generalising by syntactic category (e.g. Kaji *et al.*, 1992);
- Generalising by semantic features (e.g. Matsumoto & Kitamura, 1995);
- Generalising Parse Trees (e.g. Way, 2003);
- Generalising Strings (e.g. Cicekli & Güvenir, 2003; McTait, 2003);
- Generalising using Placeables (e.g. Brown, 1999).

Example-Based Machine Translation

All this material on alignment is a prerequisite for EBMT. To link the two explicitly, we:

- give a basic outline of the EBMT process;
- compare and contrast TM and EBMT;
- present the major problems for EBMT of *boundary definition* and *boundary friction*;
- discuss issues pertaining to the storage of examples and their segmentation, and explain the matching and recombination stages of EBMT.

Statistical Machine Translation

We present the IBM Models 1 and 2 of SMT.

Hybrid Models

Finally, given that these students have previously taken a module on RBMT, we discuss possible hybrid models.

Assessing the Students

The course was designed to be an 'assessment only' module (i.e. no end of module exam).

There were two assignments:

1. a labtest on building an aligner (week 5);
2. a group presentation/demonstration on building an EBMT system (week 8).

The labtest was a 3-hour assessment, in which the students were individually asked to develop a number of programs in Perl:

- to calculate the average sentence length of the $\langle s,t \rangle$ sentences provided in terms of both words and characters;
- to calculate the ratio of $\langle s,t \rangle$ words and characters per sentence;
- to write a length-based sentence aligner, in terms of both words and characters;
- to compare the alignment results against a 'gold standard' provided;
- to segment the 'gold standard' reference solution according to the marker hypothesis;
- to propose sub-sentential alignments using the marker hypothesis.

In addition, there were three discussion questions on aspects of the course.

Problems with Lab Test

Note that all of these programs had been tackled in the lab sessions during the course. We considered three hours to be a reasonable time limit given that we wrote programs to perform the various tasks in one hour.

However, none of the 9 students completed all questions, and in general, too many students spent far too long on programs for which very few marks were awarded (as indicated on the question paper).

Despite the fact that the students' answers were marked benevolently, over half of the class had failed, with the top mark being just 57%. However, a compromise was developed whereby the marks were divided by 0.7 in order to give a truer indication of each student's performance (top mark 81%, lowest 19%, average 53%).

Assignment 2: the EBMT System

This was a group project, where the students were divided into groups of three and were asked to develop an EBMT system, based on the sentence-, phrasal- and word-level alignments written in preparation for the first assignment. Marks were awarded both for system design and functionality, and for documentation. No one segmentation method was preferred over any other; indeed, some groups used the marker-based approach, others used a bigram approach, etc.

The groups presented their systems to these authors, who found their efforts to be extremely good (highest mark 90%, lowest 57%, average 66%).

Problems with Assignment 2

How to tell who's done what in each team?

Final Mark for Module

Finally, in order to derive the final mark for the module, the first assignment was weighted 0.35, with the group assignment weighted 0.65. All students passed the module (highest mark 87%, lowest 44%, average 65%).

Lessons Learned

- There are problems in assessment-only modules;
- We overestimated what the students could do in a 3-hour lab exam;
- When asked to develop an EBMT system in small groups, they did well.

Improvements/Changes to the Course

- The students may be assessed on a more ongoing basis in weekly labs rather than in one lab exam;
- We may use the material in (Carl & Way, 2003) in study classes, with students presenting this material to the class on a weekly basis and this constituting a percentage of their mark for the module;
- We may choose to focus more on the building of SMT systems, using the Egypt toolkit.

Concluding Remarks

- Empirical approaches to NLP and MT have reached a reasonable stage of maturity. It is important to teach these tools and techniques to University students who may become future researchers and developers in these areas;
- No (other) dedicated courses on statistical methods in MT exist;
- Almost no textbooks on empirical MT exist;
- We developed a course focussing on Alignment, EBMT and SMT;
- We commented on some of the problems we faced, and some possible improvements to the course;
- Our course would be improved by hearing about other similar courses and methods of assessment.