Teaching Statistical Machine Translation

Kevin Knight

USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey CA 90292
knight@isi.edu

Abstract

This paper describes some resources for introducing concepts of statistical machine translation. Students using these resources are not required to have any particular background in computational linguistics or mathematics.

1. Introduction

This paper describes three resources for introducing concepts of statistical machine translation. The first consists of parallel corpora and tools to support a human translation task. The second is a tutorial workbook. The third is a multiple translation corpus put out by the Linguistic Data Consortium. Working with these resources requires no particular background in computational linguistics or mathematics.

2. Human Translation Using Parallel Corpora

Computers look at parallel corpora very differently from people. To help get students into the "mind" of an automatic statistical translator, we put together a small bilingual corpus (12 sentence pairs) where the language pair is "Centauri/Arcturan" ([Knight, 1997] contains the corpus). None of the words in this corpus are understandable to humans.

Furthermore, students are given three new Centauri sentences to translate. Their only resource is the bilingual text. This problem can be solved in a few hours. Students learn:

- What parallel corpora look like.
- To view parallel corpora through the eyes of a computer.
- How parallel corpora are relevant to machine translation.
- How to build bilingual dictionaries from parallel corpora.
- How cognate information may be useful in machine translation.
- How to do word alignment, and how to employ the pigeonhole principle.
- About the chicken-and-egg nature of dictionaries (which enable word alignments) and word alignments (which enable dictionary building).

These concepts can be learned without any prior instruction – students have to learn them to solve the task. Later, it can be revealed that the "Centauri/Arcturan" corpus is really a lightly disguised Spanish/English corpus.

A somewhat larger example of the same exercise can be found at:

www.isi.edu/natural-language/mt/contest

This houses a collection of 1100 real English/Tetun sentence pairs (Tetun is a major language of East Timor), plus a monolingual Tetun news article of 10 sentences, to be translated/decoded by hand. A search tool is provided that returns all sentence pairs containing any requested monolingual word or phrase. Students learn:

- About word alignment and dictionary building at a larger scale.
- About phrase-to-phrase alignment, the norm in real translation data.
- About unalignable function words.
- The importance of knowing the target language (versus source) in making fluent translations.
- The importance of short sentence pairs (where alignment possibilities are restricted) in helping disambiguate/align longer sentence pairs.
- About locality in word order shifts.
- How to guess the meanings/translations of unknown words.
- About how much uncertainty the machine faces in working with limited data.

Tetun has non-standardized spelling (the name East Timor is spelled seven different ways in this Tetun corpus) but virtually no morphological inflection.

Students may also get ideas about machine translation algorithms after doing the job manually. [Al-Onaizan et al, 2000] describes a translation contest using this corpus, and gives results of de-briefing the winners.

3. Statistical Machine Translation Tutorial Workbook

This short workbook (www.isi.edu/~knight) gives a gentle introduction to the IBM statistical MT models, presenting the concepts in [Brown et al 93]. The presentation is mathematically from scratch, and many exercises are provided. Students learn:

- About Bayes Rule and noisy-channel probabilistic models.
- About n-gram language models and smoothing.
- About generative translation modeling, in particular Model 3 from [Brown et al 1993].
- About automatic hidden parameter estimation and word alignment via the EM algorithm.

This workbook was written in 1999 and already lacked in-depth discussion of decoding algorithms. Some significant advances in statistical MT since then include phrasal translation models, maximum entropy models, and alternatives to training using maximum likelihood. However, completing this workbook is good background for understanding this subsequent work.

4. Multiple Translation Corpus

The Linguistic Data Consortium (www.ldc.upenn.edu) has issued a very interesting data set of 100 Chinese news texts with 10 English translations each. It is called the Multiple Translation Corpus (MTC). This data has been used for automatic MT evaluation, following [Papineni et al, 2002], and also for paraphrasing research [Pang et al, 2003]. It is interesting in its own right, and by studying it, students learn:

- There is no right answer in translation, but there are wrong answers. Translators make a lot of mistakes.
- There is a lot of variation in translation, but by the ninth translation, a surprisingly large amount of that variation has already been observed.
- There are many phrases that do not admit variation, and all translators use the same wording.

This corpus is also a good jumping-off point for the algorithmic/statistical study of automatic MT evaluation.

5. References

[Brown et al, 1993] "The Mathematics of Statistical Machine Translation: Parameter Estimation", P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer. *Computational Linguistics*, 19(2).

[Knight, 1997] "Automating Knowledge Acquisition for Machine Translation", K. Knight, *AI Magazine*, 18(4).

[Al-Onaizan et al, 2000] "Translating with Scarce Resources", Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, K. Yamada, AAAI-2000.

[Pang et al, 2003] "Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences," B. Pang, K. Knight, and D. Marcu. NAACL-HLT-2003.

[Papineni et al, 2002] "Corpus-based Comprehensive and Disagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results", K. Papineni, S. Roukos, T. Ward, J. Henderson, F. Reeder. NAACL-HLT-2002.