MACHINE TRANSLATION IN THE YEAR 2004

Kevin Knight and Daniel Marcu

Information Sciences Institute and Department of Computer Science
The Viterbi School of Engineering, University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{knight,marcu}@isi.edu

ABSTRACT

Machine translation (MT) accuracy has recently increased, due to better techniques and to the availability of larger parallel training sets. Statistical MT systems are now able to translate across a wide variety of language pairs. This paper covers the basic elements of state-of-the-art, statistical MT, including modeling, decoding, evaluation, and data preparation.

1. INTRODUCTION

There are many approaches to the machine translation of human languages. Some approaches require manual knowledge entry by highly skilled linguists, while others make use of automatic training procedures. Some approaches make use of abstract meaning representations, while others work at the level of word substitution. Many combinations of these dimensions have been explored – manual entry of large dictionaries, automatic learning of phrase substitution tables, semi-automatic construction of syntactic-transformation rules, etc.

Automatic statistical training has recently made a major impact on MT accuracy. The field is evolving rapidly, as we can observe in these sample English MT outputs from the same original Arabic input document:

Best system in NIST 2002 MT evaluation:	Best system in NIST 2003 MT evaluation:
insistent Wednesday may recurred her trips to Libya tomorrow for flying	**Egyptair Has Tomorrow to Resume Its Flights to Libya**
Cairo 6-4 (AFP) - an official announced today in the Egyptian lines company for flying Tuesday is a company "insistent for flying" may resumed a consideration of a day Wednesday tomorrow her trips to Libya of Security Council decision trace international the imposed ban comment. And said the official "the institution sent a speech to Ministry of Foreign Affairs of lifting on Libya air , a situation her receiving replying are so a trip will pull to Libya a morning Wednesday".	Cairo 4-6 (AFP) - said an official at the Egyptian Aviation Company today that the company egyptair may resume as of tomorrow, Wednesday its flights to Libya after the International Security Council resolution to the suspension of the embargo imposed on Libya. "The official said that the company had sent a letter to the Ministry of Foreign Affairs, information on the lifting of the air embargo on Libya, where it had received a response, the first take off a trip to Libya on Wednesday morning".

We believe that the transition from gibberish to understandable outputs can be attributed to the increased availability of parallel data and progress in modeling, decoding, and automatic evaluation. In this paper, we review these basic elements of statistical MT, as currently practiced in the year 2004. Given the current pace of progress, it is likely that many of the techniques described in this paper will become obsolete in the near future.

2. TRAINING DATA

Statistical MT systems are trained on bilingual (human translated) documents, usually from natural sources like the French/English Canadian parliament proceedings or United Nations transcripts. Such collections have been growing in size, as demonstrated by the amounts of Chinese/English, Arabic/English, and French/English materials distributed widely to researchers over time (see Figure 1). Millions of words of bilingual data exist for dozens of language pairs. Most researchers believe that deep secrets of translation lie buried in these large data sets, waiting to be uncovered by automatic analysis.

Natural bilingual data usually needs substantial cleaning before it can be used – text de-formatting, encoding detection/conversion, and so on. It must also be

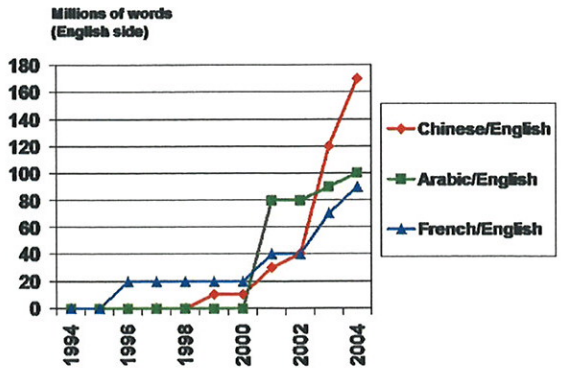

Figure 1: Data in sentence-pair format, available to researchers from the Linguistic Data Consortium.

sentence aligned, as statistical MT training procedures normally expect sentence pairs, not document pairs. Of course, sentences frequently do not match up one-for-one, and sections of text are often completely dropped by translators. To deal with such noise, and to produce shorter pairs for improved statistical training, a recent trend is to produce "segment-aligned" text, rather than sentence-aligned text, where segments may be smaller than sentences [4]. Another trend is to extract parallel segments from non-parallel corpora [11].

Text is also tokenized into words, with punctuation marks usually treated as "words" themselves. (MT researchers use the word *segmentation* instead of *tokenization* when the language is particularly hard to tokenize – e.g., Chinese or Thai). Finally, text is usually lower-cased prior to training. The resulting loss of information is considered outweighed by the positive reduction in vocabulary size (and therefore in the size of the learned translation dictionaries). MT output is typically recapitalized in a post-processing step.

3. EVALUATION

For decades, it was assumed that evaluating the quality of an MT engine is necessarily a subjective process. Two factors contributed to this state of affairs.

1. Since translation is a generation (and not a classification) task, it was unclear how one could associate a gold standard to a given test set. Table 1, which lists Chinese-to-English translations produced by 11 distinct translation agencies, shows the high degree of variability in human translated data.
2. Because the process of translation induces large scale word/phrase movements (see Table 1), it was unclear how one could measure the "distance" between machine and human produced translations.

1. At least 12 people were killed in the battle last week.
2. At least 12 people lost their lives in last week's fighting.
3. Last week's fight took at least 12 lives.
4. The fighting last week killed at least 12.
5. The battle of last week killed at least 12 persons.
6. At least 12 persons died in the fighting last week.
7. At least 12 died in the battle last week.
8. At least 12 people were killed in the fighting last week.
9. During last week's fighting, at least 12 people died.
10. Last week at least twelve people died in the fighting.
11. Last week's fighting took the lives of twelve people.

Table 1: Human translations of a Chinese sentence.

Until recently, MT researchers could not validate their ideas in a fast develop/test/evaluate cycle due to the subjective nature of the evaluation process, which employed sophisticated protocols for counting the number of lexical, syntactic, and semantic errors. However, an influential paper by Papineni et al. [15] showed that translation performance (adequacy and fluency) correlates well with the number of n-grams that co-occur in a translated document and a set of reference translations: the higher the overlap, the higher the performance of the system. Several other objective metrics, such as Word Error Rate and Position Independent Error Rate [12] also appear to correlate well with translation quality. Only two years after publication [15], the Bleu metric has become the metric of choice for measuring progress in DARPA-sponsored annual evaluations carried out at NIST and a required ingredient in any research paper. Human translations can also be scored by Bleu, typically falling around 60% (rather than 100%, due to translator variation). The best Arabic systems score 46%.

Until now, Bleu has been an important catalyst for the recent progress in the field. It is an open question whether Bleu will remain responsive to increases in automatic translation quality, especially as statistical models that impose global grammaticality play an increasing role [3].

4. MODELS

The original translation scheme proposed at IBM [1] in the late 1980s used Bayes Rule, in which one tries to choose a translation e for source sentence f that maximizes $P(e \mid f)$, or equivalently maximizes $P(e) \cdot P(f \mid e)$. The first term has almost always been captured by a smoothed n-gram language model. There are many approaches to the second term.

The IBM Models 1 through 5 approximate the translation of e into f as a word substitution/permutation process. Each of the models has a slightly different generative probabilistic story. For example, Model 3 uses a four-step process:

1. Each English word in e is either dropped, copied, duplicated, triplicated, etc. This decision is probabilistically controlled by a fertility table φ(fertility | english-word).

2. After each resulting word, a dummy word NULL is either inserted or not. This decision is controlled by a single global parameter p. NULL words are used to generate foreign-language function words with no direct English correlate.
3. Each resulting word is translated into some foreign word, as controlled by a large translation table t(foreign-word | english-word).
4. Each word in the resulting string is permuted into a possibly-different position, as controlled by a distortion table d(position-j | position-i, length-of-e, length-of-f).

Values for entries in all four tables can be estimated with the EM algorithm over a large corpus of sentence pairs, though approximations must be made, as there appears to be no polynomial dynamic-programming EM solution for Model 3. Model 1 is simpler: it drops the distortion and fertility tables, and as a consequence, admits an efficient quadratic-time EM training procedure with a convex likelihood surface. (Quadratic in the length of the particular sentence pair being analyzed).

Another popular model for P(f | e) is called "HMM" [16]. It extends Model 1 to capture distortions, and is trainable in cubic-time. HMM distortions are relative, meaning that a word tends to locate its translation near the translation of the previous word, which encourages whole groups of words to move together in translation. Another type of model is exemplified by Inversion Transduction Grammar (ITG) [17]. This joint model P(e, f) produces a bilingual binary tree with both "normal" and "inverted" nodes. Leaves of the tree are word-translation pairs. Sentence e is read off the bottom of this tree, while sentence f is read off from a tree in which the children of inverted nodes are re-ordered. This model also encourages words to move in groups, and admits a high-polynomial EM training procedure.

While the parameters from these word-substitution models can be used to drive an MT decoder (runtime translator), they are more frequently used to produce a word-aligned bilingual corpus. This corpus consists of the best (Viterbi) set of word-to-word connections for each sentence pair. Even for this purpose, the above models suffer from a serious defect – one English word might align to several foreign words, but a foreign word can only align to a single English word. For example, the Spanish word "inmobilaria" cannot align to both English words "real" and "estate". To solve this problem, one word alignment can be built in the English-to-foreign direction, another word alignment can be built in the foreign-to-English direction, and the results can be merged.

Word-aligned corpora have been effectively exploited for constructing phrase-substitution models, which have significantly outperformed word-based models in decoding. In the Alignment Templates method [13], all phrase pairs consistent with a given word alignment are collected and counted. These counts are normalized into probabilities and smoothed. Count-based smoothing can be used, but word-pair smoothing (e.g., using Model 1 and the above-mentioned t-table) is preferred, as some low-count phrase pairs are better than others.

Finally, syntactic transfer models of P(f | e) have also been proposed and implemented [5,6,8,18]. A typical parameter of such a model is P(re-order | JJ NN), i.e., what is the chance an English adjective-noun phrase is re-ordered when translated to French? For us to estimate parameter values for these models, bilingual data must be automatically parsed (in either one language or both, depending on the model). Past syntactic models have still been word-oriented, and it remains to be seen whether they can capture (and add to) large amounts of phrase-pair data. Syntactic models have many potential benefits: (1) better control over word-reordering, (2) better control over the interpretation and generation of function words, and (3) efficient, tight integration with structured language models. This third benefit stands in contrast to speech recognition -- unlike acoustic models, syntactic translation models naturally produce trees (rather than strings), which can be directly scored by tree-based language models without the need for parsing. Better language modeling may be more important for MT than for speech – unlike a speech recognizer, an MT system must carefully synthesize a fluent, never-before-uttered sentence. A glance over state-of-the-art MT output shows that most of the foreign language material is well-accounted for in automatic translation, but target-language fluency is often not there. Likewise, in the world of human translation, target-language proficiency is more highly prized than source-language proficiency.

Translation modeling research has also moved beyond the Bayes Rule decomposition of P(e | f) into P(e) * P(f | e) [12,13,14]. As a simple example, just as in speech recognition, MT performance can be improved by raising one of these factors to a constant power. It is also useful to add a third, length-bonus factor to counteract the tendency of the other two components to prefer shorter MT outputs. Many systems now contain between five and fifteen such model components, each of which gets to cast a quantitative vote for each candidate MT output considered in decoding. A model component (also called a feature function or a knowledge source) may be as complex as a language model trained on a billion-word corpus, or as simple as a binary feature that checks whether the proposed MT output correctly balances parentheses or quotes. The parameter tuning problem is difficult because the function to be optimized is not smooth and has many local optima. Off-the-shelf optimization algorithms, such as Simplex, and MT-specific optimization algorithms [12] have led to significant increases in performance over baseline systems.

5. DECODING

The most widely used decoders today are those built in conjunction with word-to-word and phrase-to-phrase translation models. Given a sentence f and a set of trained models M, the most generic decoders search for a translation e that maximizes a log-linear, weighted objective function defined over M. In the simplest case, M can contain only two models: a language model and a translation model.

Because automatic translation allows for word reordering, finding the output e that maximizes the objective function is NP-complete [9]. Two techniques are used in practice to circumvent this problem.

1. In a dynamic programming-based beam decoder [7,10,13,14], the output e is produced left-to-right, by incrementally constructing a lattice of partial translation hypotheses. Each partial hypothesis stores the last (source-phrase:target-word) pair used in translation; the next-to-last target word; a coverage-vector that makes explicit what source words have been already translated; a language and translation model score; other model scores. The most promising hypotheses are expanded left to right until a translation that has a coverage vector that subsumes the entire input sentence is found. The translation e is produced by traversing the backpointers associated with each node of the representation.
2. In a greedy, anytime decoder [7], the output e is obtained by rapidly creating a complete initial translation (by translating every source word into its most probable target equivalent, for example); and by modifying this translation afterwards locally, in a greedy, incremental manner, while the local changes lead to translations of higher score.

Decoders that exploit richer models of syntax [3,19] implement cross-lingual versions of stochastic parsing algorithms that have been proven successful in the context of syntactic parsing [2]. Research in word- and phrase-based decoding is in its teens; research in syntax-based decoding is in its infancy.

6. SPEECH AND MT

Speech and MT have been integrated to assist person-to-person communication in limited domains ("speech-to-speech translation"). It is still a challenge to support robust, spontaneous, two-way conversation. New efforts are also underway to translate captured speech into English text, for retrieval ("speech-to-text translation"). Integrated ASR-MT systems can, for example, provide automatic English captions for foreign news broadcasts. It is open whether it is profitable to jointly optimize parameters of ASR and MT systems, which until now have been separately optimized on ASR and MT training data.

7. REFERENCES

[1] P. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), 263-311, 1993.

[2] E. Charniak, S. Goldwater, and M. Johnson. Edge-Based Best-First Chart Parsing. Proceedings of EMNLP, 127-133, Granada, Spain, 1998.

[3] E. Charniak, K. Knight, and K. Yamada. Syntax-based Language Models for Machine Translation. *Proceedings of MT Summit IX*, New Orleans, 2003.

[4] Y. Deng, S. Kumar, and W. Byrne. Bitext Chunk Alignment for Statistical Machine Translation. *CSLP Tech Report*, Johns Hopkins University, 2004.

[5] J. Eisner. Learning Non-Isomorphic Tree Mappings for Machine Translation. *Proceedings of ACL*, Sapporo, Japan, 2003.

[6] M. Galley, M. Hopkins, K. Knight, and D. Marcu. What's in a Translation Rule? *Proceedings of HLT/NAACL*, Boston, 2004.

[7] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. Fast Decoding and Optimal Decoding for Machine Translation. *Artificial Intelligence*, 154 (1-2), 127-143, 2003.

[8] D. Gildea. Loosely Tree-Based Alignment for Machine Translation. *Proceedings of ACL*, Sapporo, Japan, 2003.

[9] K. Knight. Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, 25(4), 1999.

[10] P. Koehn, F.J. Och, and D. Marcu. Statistical Phrase-Based Translation. *Proceedings of HLT/NAACL*, Edmonton, 2003.

[11] D. Munteanu, A. Fraser, and D. Marcu. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. *Proceedings of HLT/NAACL*, Boston, 2004.

[12] F. Och. Minimal Error Rate Training for Statistical Machine Translation. *Proceedings of ACL*, 160-167, Sapporo, 2003.

[13] F.J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30 (4), 417-450, 2004.

[14] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A Smorgasbord of Features for Statistical Machine Translation. *Proceedings of HLT/NAACL*, Boston, 2004.

[15] K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of ACL*, 311-318, Philadelphia, 2002.

[16] S. Vogel, H. Ney, and C. Tillmann. HMM Word-Based Alignment in Statistical Translation. *Proceedings of COLING*, 836-841, Copenhagen, Denmark, 1996.

[17] D. Wu. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23 (3), 377-403, 1997.

[18] K. Yamada and K. Knight. A syntax-based Statistical Translation Model. *Proceedings of ACL*, 523-530, Toulouse, France, 2001.

[19] K. Yamada and K. Knight. A Decoder for Syntax-Based Statistical MT. *Proceedings of ACL*, Philadelphia, 2002.