English-to-English Statistical Machine Translation: Why and How

Ali Mohammad ALAWI@CSAIL.MIT.EDU
Federico Mora FEDERICO@CSAIL.MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge MA, 02139 USA

1. Introduction

Machine translation, the task of generating tools to translate or to help translate text from one natural language into another by means of a computerized system, has been the subject of intense research over the past five decades. It is one of the earliest proposed uses for the computer and, to date, one of the most dramatic failures of the field. From the beginning, computer scientists promised imminent perfect translation and have been consequently punished by grant committees for their failure to deliver (J. Pierce, 1966). In the past, techniques in machine translation followed the most popular techniques in artificial intelligence; it is not surprising, therefore, that for some time machine translation efforts were directed primarily in the construction of monolithic rule-based systems (Jordan & Benoit, 1999; D. J. Arnold, 1994). This effort in the research community continued until the turn-of-the-century, when it was abandoned with the advent of successful machine learning methods based on simple statistical models and large training sets.

The success of simple statistical methods over intricate and massive rule-based systems is ultimately due to our inability to plan for all natural language text in a set of rules. Natural language is an imprecise method of communication; it is ambiguous and largely unstructured, and is therefore the most expressive form of communication extant. Any set of rules tends to be fragile; a small set can usually yield surprisingly good results, but improvement beyond that point is suprisingly difficult (almost impossible). This is our understanding for why statistical methods are more successful in practice.

Despite this, statistical methods have their limitations. To date, state-of-the-art statistical methods have very little linguistic motivation, and, although they are able to handle complex sentences in a more robust (albeit dissatisfactory) fashion, they often make very common errors that could be fixed by very simple linguistic rules if a way could only be found to introduce such rules without introducing the notorious fragility that accompanies them; this is one popular avenue of current research. Rule-based methods, on the other hand, are generally quite heavily linguistically motivated.

Although further development of rule-based systems is considered a misplaced effort in the academic community, commercial systems are still largely based on this paradigm and are in widespread use in the industry. The poor performance of commercial machine translation systems is now a contemptuous by-word of popular culture. At the same time, quite a lot of literature exists due to the more than forty years of concentrated research in this direction. It is natural to ask, then, if some technique might be found to combine the sophisticated linguistic analysis of the rule-based systems with the robustness of the statistical systems. This is the topic of this paper.

2. Methods and Madness

Statistical machine translation systems are based on the following idea (we will make the canonical assumption that we wish to translate French sentences into English): let us assume that, when French people speak French, they are thinking of an English sentence; they perform some random, noisy procedure on the English, and spit out their French sentence. Our goal is to recover the English sentence. So, there is some distribution $P(e|f)$ and we wish find the maximum-likelihood hidden English sentence e given an observed French sentence f. We use Bayes' rule and arrive at the following:

$$\begin{aligned}\arg\max_e P(e|f) &= \arg\max_e \frac{P(f|e)P(e)}{P(f)} \\ &= \arg\max_e P(f|e)P(e),\end{aligned}$$

where we can omit the $P(f)$ since f is fixed. We call $P(f|e)$ the *translation model* and $P(e)$ the *language model*. It is this formulation that allows us to use very simple translation models and still achieve decent translations. This is very intuitive: if one wishes to translate French to English, it is better if they are native speakers of English and know a little French than if they are native speakers of French and know a little English.

There has been much work in developing both good translation models and good language models. The most sophisticated language models, in practice, never do much better than a simple trigram model, where one assumes that English is a Markov process with a history of two words. This is a recurring theme in Natural Language Processing: it's really easy to get a very competitive baseline that is nowhere close to perfect or even useful in practice.

3. Our Technique

Our technique is very simple and intuitive. We will start with the normal input to a statistical model: a large set of corresponding sentences in French and English, called a *parallel corpus*. We translate the French sentences into English using our rule-based system, then train our statistical system on the new parallel corpus of machine-generated vs human-generated English. Our translation model need not be terribly sophisticated here, obviously, as we are translating within one language and expect little reordering to be successful. It is our language model that will do the work here: fixing minor grammatical errors by considering word movement, insertion, and deletion against its "knowledge" of the English language.

4. Results

We tested our technique on the German-English EUROPARL corpus (Kohn, 2003), a collection human-generated parallel text from the European Parliament. Our rule-based system is the popular SYSTRAN, used without permission via google (Google, 2005). The statistical machine translation technique was IBM Model 2 (P. Brown, 1993) with one-dimensional gaussian alignments (Mohammad & Collins, 2005). The results are shown in Table 1.

Technique	BLEU Score
SYSTRAN	0.13
IBM2+1dG	0.20
IBM2+1dG + SYSTRAN	0.19

Table 1. Results from the individual and combined systems. Note that google does very poorly compared to the statistical model; this is almost certainly due to the general domain intended use of google and the (complex, but still) domain-specific training of the statistical models.

The vast success of the two statistical models over the rule-based model is most likely due to the fact that SYSTRAN's system is built with no domain-specific knowledge whatsoever, whereas the entirety of both statistical systems' knowledge of language is based on the EUROPARL corpus. Though the European Parliament is the medium of discussion in a wide array of domains and the origin of a variety of complex sentence structures, it is still a restricted domain and this is an advantage.

5. Future Work

Certainly a promising avenue of future research is in language models improved beyond the trigram model, particularly models that incorporate long-range measures of grammaticality. This could be used, perhaps to improve English produced by those who are not strong in the ways of the pen.

Furthermore, since we believe that the improvement was due primarily to the language model, we can also hope that a highly simplified, static translation model could replace the IBM Model 2, in effect producing better translations in restricted domains *without a parallel corpus* (whence we could not make use of statistical systems).

An interesting possible application of this kind of translation is in reproducing language evolution to, say, recast *Hamlet* in modern English. It is certainly possible that a language's development could be modeled by simply developing a sequence of language models building up to the current version and by allowing limited word movement and word changes even without a parallel corpus.

6. Conclusions

We have demonstrated a way of combining a general-purpose rule-based system with a simple statistical model to garner robustness and significant improvements in translation quality in restricted domains.

References

D. J. Arnold, *et al*. (1994). *Machine translation: An introductory guide*. Blackwells-NCC, London.

Google (2005). Google language tools. *http://google.com/*.

J. Pierce, *et al*. (1966). *Automatic language processing advisory committee*.

Jordan, B. D. P., & Benoit, J. (1999). A suvey of current paradigms in machine translation. *Advances in Computers*, *49*, 1–68.

Kohn, P. (2003). European parliament corpus. *http://www.statmt.org/europarl/*.

Mohammad, A., & Collins, M. J. (2005). Gaussian alignments in statistical translation models. *Thesis*.

P. Brown, *et al*. (1993). The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 263–311.