Research Statement

Kevin Knight

December 2011

I am interested in building large-scale artificial intelligence systems. I have organized my research around attacking large practical problems, such as human-language translation. I find that when we attack big questions, we wind up generating high-quality research topics, and it becomes more likely that technical successes will also be meaningful.

Natural Language Generation

In graduate school, I had made up my mind that knowledge acquisition was the main roadblock for AI. My first project out of school was on meaning-based machine translation. I considered three problems:

1. develop a practical meaning representation
2. translate source texts into that meaning representation
3. translate meaning representations into target texts

I focused on the natural language generation (NLG) part, that is, problems (1) and (3). Previous NLG work had focused on complete inputs for restricted domains, whereas we attacked large-scale NLG, with inputs formed from 50,000 concepts and 70 semantic relations. Two ideas were important: (1) the use of statistical models of the target language, and (2) a grammar-based refinement strategy by which semantic structures were gradually converted into syntactic ones. We represented billions of English realizations efficiently in a forest structure, and we had to solve technical problems like extracting the most likely sentence from a forest.

After my PhD student Irene Langkilde's thesis on this topic, there followed quite a bit of work at other institutions. I believe this capability is critical for ultimate success in two key NLG applications, translation and creative language generation.

- *"Two-Level, Many-Paths Generation," (K. Knight and V. Hatzivassiloglou), Proceedings of the Conference of the Association for Computational Linguistics (ACL), 1995.*
- *"Generation that Exploits Corpus-based Statistical Knowledge," (I. Langkilde and K. Knight), Proceedings of the Conference of the Association for Computational Linguistics (ACL), 1998.*

Transliteration

In building a large-scale meaning-based translator, we ran into many practical problems, such as how to deal with unknown source words. When translating from Japanese, a large fraction of unknown words are transliterations of names and technical terms into Japanese katakana script. The Japanese katakana can be romanized—e.g., *aisukuriimu* and *anjiranaito*—but it is quite difficult to reconstruct and translate into English.

With USC undergraduate Jonathan Graehl, I published the first work on how to automatically transliterate such unknown words into English. This turns out to be a very hard problem even for human translators, and in some cases, our machine algorithm was able to exceed human performance. Our technical solution made use of new weighted finite-state technology that had been developed at AT&T. Jonathan built one

of the early finite-state toolkits (USC's Carmel), which included unsupervised EM training on string input/output databases. This turned out to be a secret weapon via which we could prototype in all the areas described in this research statement.

Transliteration itself has since expanded into a substantial sub-area of natural language processing, with an extensive literature. Last year's workshop on transliteration (NEWS 2009) was the largest held at our main international conference (ACL 2009).

- *"Machine Transliteration," (K. Knight and J. Graehl), Computational Linguistics, 24(4), 1998.*
- *"Learning Phoneme Mappings for Transliteration without Parallel Data," (S. Ravi and K. Knight), Proceedings of the Conference of the North American Association for Computational Linguistics (NAACL), 2009.*

Statistical Machine Translation

I obtained my PhD before statistical methods became popular in natural language work. Because I believed that knowledge acquisition was our main roadblock, I was attracted to the methodology. I spent a long time dissecting Brown et al's (1993) seminal treatise on statistical methods for language translation, and I built new, related applications like generation and transliteration.

In 1999, I was invited to lead a statistical translation workshop at Johns Hopkins University, funded by NSF and the Department of Defense. At that intensive six-week workshop, we built and released training software modules that are still used by research groups today. The workshop allowed me to interact with excellent researchers and fantastic PhD/undergraduate students like Franz Josef Och and Noah Smith, who are leaders in the field today.

Since 2000, I have led a statistical translation group at USC, with a special focus on linguistically plausible models using natural language syntax. It seems natural that syntax should be an important component of translation models, controlling sentence re-ordering and the use of function words. However, based on previous experiences in speech recognition, many believed that such models could not work in practice. In 2006, we showed that syntax-based models were not only theoretically interesting, but that they could outperform state-of-the-art string-based models—we demonstrated this by participating in NIST-sponsored evaluations of translation quality. We were able to obtain the top score in Chinese/English translation for several years, and in 2009, we also obtained the top score for Urdu/English.

Many key ideas went into this work, including a method for extracting syntax-based rules from bilingual text (the "GHKM" algorithm), a method for binarizing syntax-based rules to accommodate target language model scoring, methods for re-structuring and re-aligning syntactic bilingual data, and so on. Syntax-based approaches to language translation have now become very popular in the research community. More generally, language translation has flourished, with hundreds of papers published per year on the topic.

- *"Fast Decoding and Optimal Decoding for Machine Translation" (U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada), Proceedings of the Conference of the Association for Computational Linguistics (ACL), 2001.*
- *"What's in a Translation Rule?" (M. Galley, M. Hopkins, K. Knight, and D. Marcu), Proceedings of the Conference of the North American Association for Computational Linguistics (NAACL-HLT), 2004.*
- *"11,001 New Features for Statistical Machine Translation," (D. Chiang, K. Knight, and W. Wang), Proceedings of the Conference of the North American Association for Computational Linguistics (NAACL), 2009.*

Tree Automata

Previous to 2004, our syntax-based language translation systems were ad hoc. Searching for a solid theoretical foundation, I studied tree automata, a class of devices invented by Bill Rounds in 1970 to formally capture Noam Chomsky's theory of transformation grammar. These automata accept and transform trees, much as finite-state machines do for strings. Tree automata were subsequently studied by European researchers for the next few decades in a pure-theory environment.

Since 2004, I worked to re-connect tree automata with natural language research. A practical outcome was a cleaner representation for language translation systems, which allowed us to build high-accuracy systems. I also developed research collaborations with European theorists like Magnus Steinby, Heiko Vogler, and Andreas Maletti. I published survey work and spoke at theory conferences in order to help bring these communities together. I also made new contributions on the theoretical side. Students and I produced new versions of abstract tree automata particularly suited to natural language processing, and we proved theorems about their expressive power, closure under composition, and so forth.

In addition, my PhD student Jonathan May built a weighted tree-automata toolkit (Tiburon) that implements intersection, composition, k-best trees, unsupervised training, projection, and many other generic operations for tree machines. This software's source code has been downloaded by hundreds of researchers and applied to problems as diverse as prosody recognition, music, and summarization.

There is now significant interest from the theory community in contributing to natural language tools and representations, as witnessed by new workshops at ACL conferences (e.g., ATANLP 2010). Exciting work is underway to study more powerful formalisms, such as synchronous tree-adjoining grammars and macro tree transducers, and how they apply to practical natural-language problems.

- *"An Overview of Probabilistic Tree Transducers for Natural Language Processing," (K. Knight, J. Graehl), Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Lecture Notes in Computer Science, Springer Verlag, 2005.*
- *"Training Tree Transducers," (J. Graehl, K. Knight, and J. May), Computational Linguistics, 34(3), 2008.*
- *"The Power of Extended Top-Down Tree Transducers," (A. Maletti, J. Graehl, M. Hopkins, and K. Knight), SIAM J. Comput., 39(2), 2009.*

Decipherment

I have recently been working on bringing classical cryptographic and natural language techniques together to solve new problems. These areas have a natural affinity—statistical natural language techniques came directly out of World War II decipherment. N-gram frequencies and Good-Turing smoothing can be traced directly back to those activities, and machine-learning techniques like expectation-maximization training likewise came out of post-war decipherment laboratories. Archaeological decipherment likewise scored successes in interpreting Mayan glyphs and Linear B. I am interested in both applications and techniques.

For example, statistical language translation requires large amounts of bilingual (parallel) training data, but such data does not exist for all language pairs and domains. If we can train on non-parallel data, many new possibilities open up. Cryptanalysts rarely have the luxury of ciphertext paired with plaintext. Instead, by correlating how symbols co-occur within ciphertext, cryptanalysts uncover hidden substitution tables and decipher their input. Likewise, if we view foreign-language text as a cipher for English—a very complex one—we can make headway.

For the past few years, students and I have been progressing through a hierarchy of decipherment problems, from letter substitution to phoneme substitution (with applications to transliteration) to word substitution, and to translation itself. Along the way, we have been able to quantify the value of non-parallel data

and to empirically test decipherability formulas presented by Claude Shannon in the 1940s. Together with Regina Barzilay of MIT, we have also begun work on deciphering ancient scripts, such as Ugaritic tablets and the medieval Voynich Manuscript. We have also applied our decipherment techniques to traditional natural language problems like part-of-speech tagging and word alignment. In particular, we have gotten very good empirical results by searching for minimal models using integer programming.

Decipherment is not yet an ACL conference submission keyword or workshop topic, but we have now presented a number of papers and tutorials. I expect this will develop into an interesting area, with applications to humanities and to the translation technology which originally inspired this work.

- *"A Computational Approach to Deciphering Unknown Scripts," (K. Knight and K. Yamada), Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, 1999.*
- *"Attacking Decipherment Problems Optimally with Low-Order N-gram Models," (S. Ravi and K. Knight), Cryptologia, 33(4), 2009.*
- *"Minimized Models for Unsupervised Part-of-Speech Tagging," (S. Ravi and K. Knight), Proceedings of the Conference of the Association for Computational Linguistics (ACL), 2009.*

New Areas

Whole-Sentence Semantics. I am now very active in creating a semantics-based machine translation research programme. New systems in this programme will explicitly represent and reason about "who did what to do" as they translate, rather than focusing on substrings and syntactic parts of speech. Part of this research involves creating a large bank of hand-built semantic structures that cover all words in the annotated sentences. Another part involves designing new automata for accepting and transforming semantic structures represented as directed graphs, extending our previous work on string automata and tree automata. I am excited about the prospect of more accurate machine translation through deeper understanding, and about tackling new challenges in converting source texts into meanings, packing alternative meanings efficiently for integrated search, and solving many other necessary linguistic and combinatorial problems.

Creative language generation. Beginning in 2009, we have developed machine algorithms to analyze large online collections of English poetry. From this analysis, we extract knowledge of how poems are put together—semantically, syntactically, and rhythmically—and we use that knowledge to generate new poems and to translate existing poems from one language to another. Of course, poetry translation is quite a challenge, but machines do have some key, exploitable advantages (what five-syllable word starts with "c" and rhymes with "ballistic"?). I am sure that the creative construction of fiction, poetry, and advertising will soon become fertile ground for computational linguistics.