MINDS Workshops

Machine Translation Working Group

Final Report

Alon Lavie, David Yarowsky, Kevin Knight, Chris Callison-Burch, Nizar Habash, Teruko Mitamura

This report is one of five reports that were based on the MINDS workshops, led by Donna Harman (NIST) and sponsored by Heather McCallum-Bayliss of the Disruptive Technology Office of the Office of the Director of National Intelligence's Office of Science and Technology (ODNI/ADDNI/S&T/DTO). To find the rest of the reports, and an executive overview, please see http://www.itl.nist.gov/iaui/894.02/minds.html.

1. Historical Perspective

The field of Machine Translation (MT) has dramatically evolved over the course of the last two decades. The dominant paradigm of the early years (starting in the 1950s, and continuing to a major extent through the 1980s) can be characterized by knowledge-rich rule-based systems, constructed via intensive amounts of labor by human experts. The large-coverage commercial systems (such as SYSTRAN [cite]), many of which are still in use to this very day, took decades of human labor to build. These systems are difficult to extend or modify, and while useful in many scenarios, have failed to achieve truly fully-automatic broad-coverage high-quality MT performance. The most significant paradigm shift in our field started evolving in the late 1980s and early 1990s. The general theme underlying this shift has been the move from manually crafted rule-based systems to a general paradigm that has *computational search* as its core. Fundamentally, this paradigm involves applying models that transform segments of input in a source language into a collection of possible corresponding target language output structures, and then executing a search for the *hypothesis* (combination of output structures) that optimizes a scoring function. The transformation models are often (but not always) acquired automatically from large volumes of sentence-aligned bilingual data. The search process is commonly referred to as *"decoding"*. A similar paradigm has been the dominant and highly successful approach used in the field of Speech Recognition, where it was formulated in the 1970s and 1980s based on the "noisy-channel model". In MT, this in turn inspired the original Statistical MT models developed at IBM in the early 1990s [BrownEtAl1990, BrownEtAl1993] and its more recent phrase-based successors [koehn-och-marcu:2003:HLTNAACL, VenugopalEffectiveACL41, chiang:2005:ACL]. The same basic paradigm also underlies various syntax-based statistical approaches to MT over the past few years [Knight et al, Wu-ITGs, etc.]. This fundamental search-based paradigm, however, has become pervasive in the field of MT, far beyond just Statistical

MT. It underlies Example-based MT approaches [brown:Coling00, Way03-cl, SatoNagao1990], modern transfer-based approaches [Lavie:TALIP:2003, imamura-EtAl:2004:COLING], a recent new approach that has been titled "Context-based Machine Translation" (CBMT) [Carbonell et al, 2006] and also approaches that combine the output from multiple MT systems (Multi-engine MT) [Lavie:MEMT:2005, nomoto:2004:ACL]. This major paradigm shift in MT was enabled by a combination of several important developments in Computational Linguistics and computing technology. Perhaps the most important of these enabling factors have been (1) the increasing availability of vast amounts of monolingual and bilingual (sentence-parallel) text corpora; and (2) the common (and increasingly inexpensive) access to fast computing power that supports tractable execution of search algorithms within increasingly large hypothesis spaces. These two important factors have also inspired similar paradigm shifts in our sister fields of Speech Recognition, Information Retrieval and NLP over the past two decades. They have also been crucially important to major developments in the field of Machine Learning, along with significant theoretical advances in that field. Machine Learning in turn, is increasingly influencing the types of models used in MT and the underlying algorithms used in automatic learning of these models from parallel bilingual data. Another important factor in the case of MT has been the advent of automatic metrics for MT evaluation. While the metrics developed so far are not accurate enough to fully replace human evaluation of translation quality, they provide rough targets for ongoing system testing and development, resolving the "bottleneck" of dependence on human assessment. Moreover, they provide target functions for parameter optimization processes that are central to state-of-the-art search-based MT approaches.

2. Weaknesses and Limitations in Current State-of-the-Art MT

While search-based MT has become the dominant paradigm over the past decade, and research advances have been both significant and impressive, the ultimate goal of fully-automatic broad-coverage high-quality MT for many language pairs remains beyond current state-of-the-art. The highest levels of translation quality and fluency are currently achieved by phrase-based statistical MT systems that are trained using very large volumes of sentence-parallel text. Such volumes of parallel text exist for only a small number of language pairs (the largest parallel corpora covering mostly the major European languages, Arabic/English, Chinese/English and Japanese/English). These large parallel corpora are predominantly in limited language genres and styles such as Newswire and Parliamentary Proceedings. The MT systems that result from training on such data often produce reasonable-quality translations for sentences similar in vocabulary and style to the training data, but deteriorate quickly as the input style or genre strays from that seen in the training data. Thus, not only are today's high-quality MT systems limited to a handful of language pairs, but they are further limited by text genre and style.

We believe that current state-of-the-art search-based MT approaches can be characterized by the following main common fundamental weaknesses:

1. **Weak Models:** The models used by current MT approaches are not strong enough to consistently generate correct translations. Consequently, the hypothesis-spaces that are generated by these MT approaches often do not contain correct, or even good

possible translations of the input. The fragility of today's phrase-based statistical MT systems, in particular, is a direct consequence of the fact that the models underlying these systems are both linguistically shallow and highly lexicalized. Fundamentally, they only "learn" likely phrase-to-phrase correspondences between source and target languages, by observing such correspondences in large amounts of raw parallel data. These models are attractive because they are simple and can be applied to un-annotated naturally occurring data, but they do not capture language correspondences at levels that are useful for adequate generalization, and they are not capable of enforcing grammatical target language output. Similar weaknesses can be observed in other search-based approaches such as EBMT and Multi-Engine MT. While there is fairly broad recognition in the research community that translation correspondences can be more effectively described using representations that explicitly model syntactic divergences between languages (such as synchronous context-free grammars) and differences in syntax-to-semantic mappings (such as dependency grammar transducers), how to effectively acquire such models from parallel data is a daunting open problem. More recent work on syntax-based and factored statistical translation models [cites] hope to at least partially address these issues.

2. **Weak Discrimination During Search:** The knowledge resources utilized in today's MT systems are insufficient for effectively discriminating between good translations and bad translations. Consequently, the decoders used in these MT systems are not very effective in identifying and selecting good translations even when these translations are present in the search space. The most dominant knowledge source in today's decoders is a target language model (LM). The language models used by most if not all of today's state-of-the-art MT systems are traditional statistical n-gram models. These LMs were originally developed within the speech recognition research community. MT researchers later adopted these LMs for their systems, often "as is." Recent work has shown that statistical trigram LMs are often too weak to effectively distinguish between more fluent grammatical translations and their poor alternatives. Numerous studies, involving a variety of different types of search-based MT systems have demonstrated that the search space explored by the MT system in fact contains translation hypotheses that are of significantly better quality than the ones that are selected by current decoders (see [cite] for one example), but the scoring functions used during decoding are not capable of identifying these good translations. Recently, MT research groups have been moving to longer n-gram statistical LMs, but estimating the probabilities with these LMs requires vast computational resources. Google, for example, uses an immense distributed computer farm to work with 6-gram LMs. These new LM approaches have resulted in small improvements in MT quality, but have not fundamentally solved the problem. There is a dire need for developing novel approaches to language modeling, specific to the unique characteristics of MT, and that can provide significantly improved discrimination between "better" and "worse" translation hypotheses.

3. Recommendations for New Major Research Themes

In the course of our group discussions at the two MINDS workshops, a broad range of ideas and suggestions for new themes and research directions were proposed and analyzed. In the recommendations that follow, we first identify three "grand challenges"

- major capability goals that we believe should drive MT research priorities in the future. We then identify the main technical advances in MT that we feel are necessary in order to push the field towards reaching these three main goals. Many of these technical advances contribute towards all three goals, but some are unique to just one of the goals. We then elaborate on five specific technical research themes, which we feel have the potential to have the greatest impact on MT progress towards achieving the desired improved MT capabilities. It is our belief that the three most important capability goals that MT research should strive for are: **(1) High-Quality MT for many more language pairs; (2) Substantially improved translation quality robustness across domains, genres and language styles;** and **(3) Achieving human-level translation quality and fluency.**

Required Technical Advances for MT for Many Languages:

- **MT Models Trainable with Limited Data Resources:** Over the past decade, MT research in general , and statistical MT in particular, has focused on a small number of language pairs for which vast amounts of sentence-aligned parallel text have become available (or was explicitly constructed). While parallel corpora are becoming increasingly available for additional language pairs, the magnitude of such corpora that is likely to be available for most language pairs in the foreseeable future is limited. The levels of translation performance that can be achieved using today's MT models with such limited amounts of data are rather unsatisfying. Significant progress is therefore required in developing new types of translation models that better generalize from limited amounts of available training data, in order to enable the development of MT systems for a far broader range of language pairs.
- **MT Models Suitable for Morphologically-complex Languages:** Much of the MT research work over the past decade has focused on translation from a small number of languages (predominantly Arabic and Chinese) *into English.* While Arabic has a somewhat complex morphology, English morphology is comparatively quite simple, and Chinese has very little word-level morphology. Consequently, recent MT research work has placed only limited focus on issues of effectively handling languages with complex morphology, and especially *generating into* target languages with complex morphology. Data-driven MT approaches to date have often simply ignored morphology and trained on fully inflected word forms. Some recent work on factored models and syntax-based approaches has begun addressing these issues, but are at very early stages. As many of the languages of the word have complex morphology, MT to and from these languages will require significant progress on developing new translation models that can effectively deal with complex morphology. The expected limited amounts of parallel text for such languages greatly amplifies the magnitude of this problem, as training MT models on full-formed words that naturally occur in the data will clearly suffer from extreme data sparseness issues.
- **Effective Language Models for Many Languages:** Language models are critical components in all search-based MT approaches. The statistical language modeling methods employed in today's MT systems have primarily focused on English. While these English LMs are only partly effective, the situation is even more challenging for other target languages, especially ones that are

morphologically more complex. Not only does the morphological complexity require larger amounts of monolingual training data in order to overcome issues of data sparseness in n-gram probability estimation, but the standard short-range n-gram models are not capable of enforcing longer-range morphologically-marked dependencies that are common for such languages, such as verb-argument agreement markings and/or case markings. MT into such morphologically-rich languages will require novel approaches to language modeling, which can better capture these types of constraints, and discriminate between more grammatical MT hypotheses and their poorer alternatives.

Required Technical Advances for Domain and Genre Robustness:

- **MT Models that Generalize Well:** As noted above, one characteristic of current phrase-based statistical MT approaches is that they are completely lexicalized. These translation models fail to capture general syntactic patterns that reflect how syntactic structures in the source language should map to their corresponding structures in the target language. The resulting MT systems are extremely brittle to changes in vocabulary and text style, which is common when systems trained on text from one domain or genre are applied to text from a new domain or in a different genre. Overcoming this "robustness" problem requires developing translation models that generalize far better from their given training data. For example, the most common general patterns of mappings between basic syntactic structures in two languages are likely to hold across genes and text styles. Such patterns can be acquired from training data in one domain, and yet be effective when translating text from a different domain. Word-level translation pairs for the new domain still must be acquired somehow, but this can conceivably be done using far smaller amounts of new domain-specific training data.
- **Model Adaptation to new Genres and Domains:** The types of models employed by state-of-the-art MT approaches as well as the models we envision in the future are statistical in nature. The probabilities of these models are estimated from training data using a variety of estimation techniques. When porting such systems into new genres and/or domains, the underlying model probability space changes significantly. Adaptation processes need to then be applied to the problem of re-estimating model probabilities. While there has been some work on how to effectively perform model adaptation in recent years, we feel that much more progress is needed in this area. As the field develops advanced new types of models, new adaptation techniques, most suitable for these new models will need to be developed and explored.
- **Targeted and Active Learning from Parallel Data:** In many scenarios, ample amounts of data for training MT systems will be available in specific domains and text styles. Only very limited amounts of training data may be available in targeted genres and domains. A major challenge when adapting or extending MT systems to such new domains and genres is how to utilize the most out of the limited new data that is available. Very little research has been done to date on methods that can identify the differences between genres and domains and use this information for targeted learning of new models. This may involve identifying lexical coverage gaps, differences in syntactic structure or other

language characteristics. Furthermore, in some scenarios, it may be possible to actively create small amounts of targeted new training data that are most useful for improving MT system performance. Machine Learning techniques such as "Active Learning" can be explored as general frameworks for innovative solutions to this problem. Along with such active learning methods, there is a need for developing novel approaches for automatically acquiring the data resources that are identified to be most useful for adaptation. This involves a targeted search of the web for data resources (parallel sentences, monolingual text, translations for named entities, etc.) that are most useful for the task at hand.

Required Technical Advances for Human-level MT Quality:

- **MT Models based on Advanced-levels of Representation of Syntax and Semantics:** The models employed by today's MT systems are too simple and naïve for capturing many complex divergences between languages, and there is broad recognition that advancing MT performance to near-human performance levels will require translation models that can capture advanced syntax and semantic representations and how they correspond across languages. The key technical challenge is identifying representation formalisms that are rich and powerful enough on the one hand, yet are simple enough to support the development of algorithms for automatic acquisition of the models from training data, appropriately annotated. Automatically acquired syntax-based models for MT have started to receive increasing attention in the last five years. Significant progress on developing these models will require building on large-scale advances in NLP, in particular in the areas of syntactic and semantic parsing. We believe that the most promising direction for developing learning approaches for MT models based on syntax and semantics is to learn them from sentence-parallel corpora that are accurately annotated with syntax and semantic structures. The main learning task in this case is to discover the correspondences between the structures in the two languages and to model these correspondences statistically. A more challenging scenario is to learn such models using parallel data where syntax and semantic structures are available for only one of the two "sides". The learning task in this case is to project the structures from one language to their corresponding structures in the other language, using crude and "noisy" information about word-to-word correspondences. The bottleneck is obtaining the needed training resources. Only small and very limited annotated corpora of this kind currently exist. The development of such annotated corpora is a critical enabling step, without which this research direction cannot hope to even get started.
- **MT Models that Incorporate Inter-sentential Context:** Practically all MT work to date has focused on translation at the level of individual sentences. Sentences form a natural "unit" of translation, as the meaning expressed in a single sentence in one language can commonly be expressed in an appropriately translated sentence in the other language. This observation is the underlying key reason why sentence-parallel corpora serve as the predominant training data for MT. Nevertheless, human-quality text in any language clearly has discourse structure, which is used to express various aspects of meaning, and such discourse

structure may also differ across languages. Furthermore, the correct translation of pronominal referents and other entities often requires resolving the references using inter-sentential context. We therefore believe that truly human-quality MT will require approaches that can explicitly use inter-sentential information in order to resolve specific types of ambiguity and which can generate coherent multi-sentence discourse structure in the target language.

- **Semantic Pivot and Statistical Interlingua Approaches to MT:** Deep-semantics and Interlingua approaches to MT received broad attention in the "old days" of MT, and have fallen out of favor in the last 15 years, because of their weaknesses that have so far appeared to be insurmountable. Automatically analyzing source language sentences all the way into a language-independent representation of meaning requires a complex series of NLP components. Current NLP technology can reliably and accurately perform this task only for limited domains, and using extremely labor-intensive coding and development by experts. Generating target sentences from Interlingua is similarly challenging. Furthermore, researchers have broadly recognized the extreme challenge of devising a true interlingua representation, that is simultaneously adequate for all languages it is intended for, rich enough to represent all intended meaning, and simple enough for humans to agree upon and for NLP algorithms to analyze and generate from. Nevertheless, we believe that research efforts on automatically acquiring complex models that can analyze and generate from interlingua representations is a worthwhile research endeavor, that carry the promise of ultimately leading to truly human-quality MT.

Proposed Concrete Research Themes

We now elaborate on five concrete research themes, which we believe have the greatest promise to create real "breakthroughs" in MT capabilities over the next five to ten years. We believe these themes can and should serve as an excellent basis for a research agenda for new funding programs in MT.

1. **Effective Sub-sentential MT Models that Generalize:** We propose to create research scenarios that explicitly encourage MT researchers to focus on the problem of developing MT models at the sub-sentential level that capture correspondences at increasing levels of syntax and semantics. As noted above, a first critical enabling step in this direction would be to create sentence-parallel corpora annotated with accurate syntactic/semantic structures on one side or both sides. Research scenarios that involve translating a broad mix of genres, domains and text styles will also encourage development of more general models, as the current models will not perform well in such scenarios. Additional research scenarios that can encourage such development include an explicit focus on specific sub-problems of MT, such as translation of sentences that exhibit known types of language divergences, which cannot adequately be handled using today's MT approaches.
2. **Learning More from Less Data:** Recent MT funding programs have spent significant effort on acquiring very large amounts of raw training data, which is clearly useful for today's phrase-based statistical MT approaches, which benefit

from such vast amounts of available data. We believe, however, that research scenarios that require learning "as much as possible" from limited amounts of data that are annotated with higher-levels of representation will create needed incentives for researchers to develop MT models that are far more suitable for realistic scenarios, where only limited amounts of data will be available. The challenge should be how to develop the best performing MT system given such limited amount of training data, and additional NLP tools and resources.

3. **MT from English into Other Languages:** As noted earlier, research programs to date have predominantly focused on MT from a small number of languages into English. While we recognize that translation into English is of primary importance to the US government in general and to the defense and intelligence agencies in particular, we feel that research scenarios that involve translation from English into other languages are of extreme importance in that they create challenges that will force MT researchers to deal with language phenomena that have not received adequate attention to date. Adequate treatment of complex morphology, described earlier, is one clear research challenge that translation into languages other than English will require.
4. **Multi-Engine Machine Translation:** The MT research community has pursued a range of different approaches to the problem over the course of time. While the field has clearly consolidated around search-based paradigms in the past ten years, we believe that different approaches using different types of models should continue to be developed, and that research scenarios that explicitly encourage such diversity in approach is in the best interest of advancing the MT field. Furthermore, there has recently been a surge in interest in approaches that can synthetically combine different MT engines operating on a common input into a "consensus" translation which surpasses all the individual MT engines in its quality. Such approaches are known as "Multi-Engine MT". We believe research into multi-engine MT is extremely important, as it is unlikely that one uniform approach to MT will outperform all others in a variety of conditions in the foreseeable future. Moreover, research scenarios where teams are evaluated not only based on the performance of their MT system in isolation, but rather in the contribution of their MT system within multi-engine combinations, will encourage research into diverse MT approaches. We strongly recommend the creation of such research scenarios in future finding programs, along with explicit funding for developing multi-engine MT.
5. **MT Evaluation:** Automatic metrics for MT evaluation have been receiving increasing attention over the past five years. Such metrics are critical tools for current and future MT research, as they allow research teams to guide the development of their systems based on frequent concrete performance evaluations. Even more critically, the models used by MT systems today and in the future contain a variety of parameters that need to be tuned for optimal performance. Automatic MT evaluation metrics such as BLEU provide the "target function" for optimizing these parameters for best translation performance. The automatic metric available today, however, are very crude, and do not correlate well with human judgments of translation quality. As MT systems improve and achieve high levels of translation quality, it becomes ever more

important to have evaluation metrics that are sensitive to small differences between translations at the sentence-level, so that minor improvements can still be detected, concrete translation errors can be isolated and identified, and system parameters can be optimized to truly achieve the best translation performance. We recommend treating MT evaluation as a research topic in its own right within future funding programs. Advances in this sub-field of MT are critical enabling steps for improving MT in general.