Preface

INTRODUCTION

Applied Natural Language Processing (ANLP) is a home to computational, cognitive, and linguistic researchers concerned with computational approaches to real-life language-related issues. More specifically, ANLP is a field predominantly interested in research that increases the ability to mimic human intelligence and human behavior, knowledge of how the mind represents and retrieves knowledge, and the ability to assess and describe how language impacts the world and the individuals and groups that comprise it.

In the first of the two books on ANLP compiled by the editors, *Applied Natural Language Processing: Identification, Investigation, Resolution*, they focused on establishing what ANLP is, where ANLP comes from, and how ANLP works. In this second book on ANLP, *Cross Disciplinary Advances in Applied Natural Language Processing: Approaches and Issues*, they provide research that is expanding the borders of ANLP from its computational linguistics birthplace into fields of research that include cognitive science, applied linguistics, corpus linguistics, and affective computing (among others). For all instances, the contributors raise applicational issues and approaches; and to this end, the first section of the current book features chapters on (1) the roles of collecting, organizing, and applying data as corpora; (2) the challenges of understanding, assessing, and reasoning with text, from both cognitive and computational stand-points; and (3) the application of such research to intelligent tutoring systems. In section 2, the editors provide readers with many examples of the best and most cutting edge ANLP research currently being conducted.

THE FIELD TRIUMVIRATE

In the first book, the editors argued that ANLP was an "emerging field," and if it were to fully emerge, then it would need to form a recognized identity. They argued that three of the major players that will help to form this identity are the fields of computational science, linguistics, and cognitive science. The role of the first two fields should seem fairly obvious, and in this book, these roles are discussed further (e.g., computing issues such as open source NLP tools are discussed in Chapter 2; and issues of natural language generation are discussed in Chapter 11; whereas linguistic issues of corpora and discussed in Chapters 4, 5, and 6; and textual analysis tools for language learning are discussed in Chapter 18). But ANLP is not just about computing and language, it is also about cognition, and therefore it is also about the field of cognitive science. Of course, the role of cognitive science isn't as obvious for all researchers, so this book (in general) and this preface (in particular) serves as a useful opportunity to describe the role, going forward, of cognitive science in ANLP.

Of course, researchers from the field of cognitive science featured prominently in the first book on ANLP. Chapters were contributed by scientists as well established as Walter Kintsch, Art Graesser, Danielle McNamara, and Jamie Pennebaker. It is researchers such as these that have helped design and develop intelligent tutoring systems, semantic analyzers, and textual analysis tools that have woven cognitive theory with linguistic features into computational approaches that have revolutionized textual assessment. Indeed, without researchers such as these, there would be no ANLP as we know it today. But this having been said, it's important to distinguish between the obvious and invaluable role of cognitive scientists on the one hand, and the less obvious role of the field of cognitive science itself. The editors hope this preface serves to somewhat highlight this distinction such that developments in ANLP can continue apace.

Cognitive science is a field dedicated to the study of how the mind works. Originally conceived as a *big tent* of highly inter-disciplinary psychologists, anthropologists, computer scientists, engineers, and linguists, cognitive science has slipped into being arguably little more than a case of *cognitive psychologists et al*. Given that "the mind" is the centerpiece of cognitive science, the demise of the other fields' participation was (perhaps) inevitable. But despite the falling away of substantial involvement from fields outside psychology, cognitive science has still managed to grow into being one of the most dominant forces in academia.

Such a position of strength for cognitive science might lead us to ask what the field has to gain from ANLP. After all, cognitive science appears to have become a thriving discipline on its own, without the need for strong ties to ANLP. To be sure, cognitive science doesn't need ANLP, but ANLP is certainly richer (and in many ways 'completed') only when the role of cognition is accommodated. As such, it is incumbent upon those interested in linguistic and computational research at an applicational level to approach and collaborate with cognitive scientists.

To understand better the need for ANLP to engage ever more closely with cognitive science, consider what constitutes success in an ANLP task. Formally, it can be said that success in a given ANLP task is the accuracy with which a computational approach can extract *illocutionary* and/or *perlocutionary acts* based overwhelmingly on nothing more than a *locutionary* form. That is, the goal of an ANLP task is to understand what is *meant* (illocutionary) and/or what effect and response that meaning has (perlocutionary) based on a string of characters typed in the text (locutionary). Parsing out this task, it is shown that filling the gap between a locutionary string and what that string actually means is primarily a linguistic task, but filling the gap between the illocutionary intent and the enactment of a response is primarily a cognitive task.

Completing the field triumvirate is computational science. It is through this field that each of the above tasks is made manifest by a process of experimentation, modeling, and implementation such that an algorithm can be derived and deployed that most reliably reflects the goals of the task at hand. As such, for an ANLP task to be successful, linguistic, computational, and also cognitive considerations have to be identified, investigated, and resolved. Thus, it is necessary that ANLP advances a cross disciplinary approach to issues (hence the name of this second volume).

ORGANIZATION OF THE BOOK

Section 1 features 10 chapters (loosely) organized around the three major fields of ANLP: computer science, linguistics, and cognitive science. Given that ANLP is nothing if not the attempt to blur such categorizations, such an arrangement may seem counter-productive; however, we must also recognize that ANLP is where *research* ends, and not necessarily where research*ers* begin. As such, the organization may serve to attract readers to an amenable point of embarkation.

Beginning with issues more closely connected with computer science, Yorick Wilks (Chapter 1) argues that a great deal of work in formal Computational Semantics (Compsem) actually includes no computation at all. Because such works lack implementation and validation, he argues that their value to NLP and Artificial Intelligence research is diminished. The author concludes that concrete computational tasks on a large scale that involve meaning representation are of primary value of Compsem. In Chapter 2, Justin Brunelle and Chutima Boonthum-Denecke give an overview of various available open source NLP tools. They discuss a subset of tools available for researchers and enthusiasts of computer science, computational linguistics, and other fields that may utilize or benefit from natural language processing. In Chapter 3, Marie-Francine Moens discusses information extraction. The discussion includes common information extraction tasks and current issues. Among these issues is the need to develop technologies that require a minimum of human supervision, to build systems that automatically acquire world knowledge, and to integrate such outputs into advanced information extraction systems.

Turning to issues more closely connected to linguistics, Charles Hall (Chapter 4) provides an overview of the history and development of the corpus, including terminology and criteria that define the modern corpus. The chapter ends with a discussion of the most basic analytical tool for corpus linguistics, the concordancer. In Chapter 5, Scott Jarvis draws attention to some of the prominent areas of overlap between Applied Linguistics and ANLP. He highlights the problems these two disciplines face in relation to the characterization of lexical deployment, focusing particularly on challenges related to the measurement of lexical diversity and the representation of the unique lexical signatures of individual samples of natural language use. In Chapter 6, Philip McCarthy and Danielle McNamara describe the User-Language Paraphrase Corpus, a freely available set of data designed to function as a challenge for researchers interested in creating or testing approaches to paraphrase evaluations. The term *user-Language* refers to the natural language input of users interacting with an intelligent tutoring system (ITS). The term *paraphrase* refers to ITS users' attempt to restate a given *target sentence* in their own words such that a produced sentence, or *user response*, has the same meaning as the target sentence. The challenge posed for researchers is to describe and assess their own approach (computational or statistical) to evaluating, characterizing, and/or categorizing, any, some, or all of the paraphrase dimensions in this corpus. In Chapter 7, Amber Chauncey Strain and Lucille Booker discuss Amazon's Mechanical Turk (MTurk) and its role in ANLP research. Amazon's Mechanical Turk (MTurk) is a Web-based data collection tool that has become a premier resource for researchers who are interested in optimizing their sample sizes and minimizing costs. The authors describe MTurk, address the issue of institutional review board processes, and discuss the benefits and limitations of using MTurk in ANLP research. In Chapter 8, Eduardo Blanco and Daniel Moldovan explore the problems of detecting negation in texts. Negation has complex interactions with other aspects of language. Thus, Blanco and Moldovan detail the forms that negation takes, and some heuristics for discovering negation automatically.

The third element of the ANLP triumvirate, cognitive science, begins with Slava Kalyuga (Chapter 9) describing cognitive load theory. Cognitive load theory is concerned with instructional consequences of the processing limitations of human cognitive systems. Because of these limitations, text processing may result in an excessive cognitive load that can influence comprehension as well as change learner affective states. The chapter reviews basic assumptions of cognitive load theory, their consequences for optimizing the design of information presentations, and implications for processing written and spoken texts. In Chapter 10, Anne Britt, Katja Wiemer, Keith Millis, Joseph Magliano, and Patty Wallace present applications to help students assess and improve their ability to reason with texts. The applications include assessing reading comprehension strategies (RSAT), enhancing scientific reasoning (CT Tutor and Operation ARIES!), teaching appropriate sourcing and integration skills (SAIF), and improving argument comprehension and evaluation skills (CASE). All of these applications deploy semantic algorithms on verbal input to assess students' performance. The goal is to provide effective assessment and feedback for learning.

In Section 2, the editors again (loosely) organize chapters around the three major fields of ANLP. However, because section 2 features ongoing research, the blur between the categories is reassuringly present. Thus, while section 1 may serve as a point of embarkation, section 2 may be viewed as a destination that researchers may be inclined to visit, or even one day come to call home.

Beginning again with subjects more closely connected to computer science, Andrew Olney, Natalie Person, and Art Graesser (Chapter 11) discuss Guru, a conversational expert ITS. Guru is designed to mimic expert human tutors using advanced applied natural language processing techniques including natural language understanding, knowledge representation, and natural language generation. In Chapter 12, Anne Kao, Stephen Poteet, David Jones, and David Augustine describe Boeing's Part Name Matching by Analysis of Text Characters (P-MATCH) system. P-MATCH is used to identify the names of parts in maintenance logs, which are often noisy, caused by employees using different spellings or abbreviations. P-MATCH is used to illustrate the value of combining natural language processing and text mining. In Chapter 13, Olga Uryupina, Massimo Poesio, Claudio Giuliano, and Kateryna Tymoshenko investigate two publicly available Web knowledge bases, Wikipedia and Yago, in an attempt to leverage semantic information and increase the performance level of a state-of-the-art coreference resolution engine. The authors propose that using disambiguation tools for Wikipedia, and adding constraints to Yago, yield the best results.

Remaining computational, but moving more clearly towards linguistic issues, Philip McCarthy, David Dufty, Christian Hempelmann, Zhiqiang Cai, Danielle McNamara and Arthur Graesser (Chapter 14) address the problem of identifying new versus given information within a text. The authors discuss a variety of computational new/given systems and analyze four typical expository and narrative texts against a widely accepted theory of new/given. In Chapter 15, William Yang Wang, Ron Artstein, Anton Leuski, and David Traum present a method for incorporating phonetic features of words into speech recognition software. By augmenting word strings with phonetic features derived from a dictionary, the authors observed a reduction in errors, with the best performance resulting from models that incorporate both word and phone features. In Chapter 16, Aqil Azmi and Nawaf AlBadia report on a method that automatically extracts and graphs the names of narrators from hadith texts (narrations originating from the words and deeds of Prophet Muhammad). The task is complex because each hadith has its own way of listing narrators, and the text of a hadith is in Arabic, a language rich in morphology. In Chapter 17, Simon Delamarre and Maryvonne Abraham present some modules from their "pictographic translator" application. The pictographic translator is an application that performs syntactical analysis of sentences directly written by the user in natural language, and then dynamically displays a series of pictograms that illustrate the words and structure of the user's sentences. The authors conclude with a discussion of the potential and limitations of the architecture of this application. The linguistics section of Section 2 ends with the most applied chapter: Rachel Rufenacht, Philip McCarthy, and Travis Lamkin (Chapter 18) investigate the potential of using traditional fairy tales for English language learners. Using the computational textual analysis software, the Gramulator, they analyze the linguistic features of fairy tales relative to a corpus of English language learners' reading material, and a corpus of baseline educational texts for native English speakers. The results show that fairy tales have the potential to be used in language learning environments.

Incorporating computation and linguistics, but with a clearer focus on cognitive science, Sidney D'Mello and Arthur Graesser (Chapter 19) survey the existing literature on text-based affect sensing. Using data from interactions with AutoTutor, an intelligent tutoring system that uses conversational dialogues, the researchers focus on how learners' affective states (boredom, flow/engagement, confusion, and frustration) can be automatically predicted by variations in the cohesiveness of tutorial dialogues. In Chapter 20, Michele I. Feist and Dedre Gentner investigate the semantics of spatial locatives. They present four studies examining the ways in which three classes of attributes – geometric, functional, and qualitative physical – influence speakers' uses of the English spatial prepositions in and on. The experiments show that all three kinds of factors play roles in English speakers' choice between these prepositions. Hansen Schwartz and Fernando Gomez (Chapter 21) present an evaluation of WordNet-based semantic similarity and relatedness measures in tasks focused on concept similarity. Concept similarity is studied and is used in many disciplines. This evaluation focuses on the application to Natural Language Processing itself. Past studies have either focused entirely on relatedness or only evaluated judgments over words rather than concepts. Eduardo Blanco, Hakki Cankaya, and Dan Moldovan (Chapter 22) describe methods for extracting *commonsense knowledge*. Commonsense knowledge encompasses facts that people know but do not communicate most of the time. That is, commonsense knowledge is the type of knowledge that machines cannot infer without being taught how (i.e. that *a person needs soap and water to shower*). The authors identify *commonsense* rules and combine those rules with a basic semantic representation in order to infer *commonsense facts*. The authors' results show that this method is able to successfully extract commonsense knowledge with high accuracy and little human interaction. And in the final chapter, Ekawat Chaowicharat and Kanlaya Naruedomkul (Chapter 23) present Co-occurrence-Based Error Correction (CBEC), a solution to the problem of word segmentation in Thai. CBEC is designed to provide accurate segmentation results based on context and purpose. CBEC quickly segments the input string using any available algorithm. Next, CBEC checks its segmentation output against an error risk data bank to determine if there is any error risk. Then, CBEC re-segments the input string using the co-occurrence score of the word sequence to ensure the accuracy of the segmentation result.

GOING FORWARD

In this preface, and in this book, the editors have attempted to demonstrate the importance of the three major contributors to ANLP: computer science, linguistics, and cognitive science. Hopefully, they have shown that ANLP stems from researchers of various homes, but culminates in research of a common interest. Of course, not all ANLP work can, will, or should equally include the three fields described here, nor can it be expected that every corner of every field will have an interest in ANLP. But all ANLP work should be of interest to its major contributors, and whereof research does not concern itself with computational solutions to real world language-related issues, thereof the research ceases to be ANLP.

With such considerations in mind, this preface may conclude how it began: by considering the role of cognitive science in ANLP, and more particularly, with some suggestions for future research that involve it. As discussed, cognitive science is an essential element of ANLP because real world language related issues are invariably human issues. But clearly, some elements of cognitive science more readily lend themselves to ANLP than others. Thus, going forward, it is likely that cognitive science moves in stages, rather than *en masse*. Of course, this much has already been seen with discourse scientists, semantic analysts, and ITS designers leading the march. But as this book hopefully demonstrates, the next great wave of cognitive science interest will be from those in affective research. That is, whatever the computational power or the linguistic parsing available in the approach, the solution to understanding and assessing text must include an acknowledgement of the cognition-emotion link (Izard, 2009; Meyer & Turner, 2006), such as the writer's levels of valence and arousal, discrete emotional states like boredom, frustration, or engagement (D'Mello & Mills, in review), anger or sarcasm (Angesleva, Reynolds, & O'Modhrain; 2004; Wilson, Wiebe, &Hwa, 2004), deceit (Duran, Hall, McCarthy, & McNamara, 2010), appraisal (Clore & Ortony, 2010; Scherer, 2001; 2005), and cognitive flexibility and decision making (Bower & Forgas, 2000; Damasio, 1995; Isen, 2010). Also, affect is likely to be qualitatively dependent on such variables as gender and age (i.e., men and women are different; the old and the young are different), and affect may differ depending on the context, theme, and topic of discourse. Further, while language is the most prominent method of symbolizing emotions across individuals and cultures, a certain level of *emotion knowledge* is needed in order for a human or artificially intelligent agent to understand the illocutionary and perlocutionary intent of locutionary acts. If ANLP researchers strive to understand the person behind the text, it is necessary to explore how these cognition-emotion links are conveyed in various types of written form.

The final section of the first book's introduction began with the claim "the future is bright." And indeed, barely a month after publication, our second book was going to press. As such, it doesn't take a mathematician to calculate that the interest in this field is substantial. Hopefully, through these two books, the editors have demonstrated the breadth of that interest, but hopefully they have also provided researchers with issues to consider, approaches to apply, and an appreciation of the cross disciplinary advances being made in the field of applied natural language processing.

Philip M. McCarthy
University of Memphis, USA & Decooda.com, USA

Travis A. Lamkin
University of Memphis, USA

Amber Chauncey Strain
University of Memphis, USA

Chutima Boonthum-Denecke
Hampton University, USA

REFERENCES

Angesleva, J., Reynolds, C., & O'Modhrain, S. (2004). EmoteMail. *Proceedings of the 31st International Conference on Computer Graphics and Interactive Techniques*. DOI: 10.1145/1186415.1186426

Bower, G. H., & Forgas, J. P. (2000). Affect, memory, and social cognition . In Eich, E., Kihlstrom, J. F., Bower, G. H., Forgas, J. P., & Niedenthal, P. M. (Eds.), *Cognition and emotion* (pp. 87–168). New York, NY: Oxford University Press.

Clore, G. L., & Ortony, A. (2010). Appraisal theories: How cognition shapes affect into emotion . In Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. (Eds.), *Handbook of emotions* (3rd ed., pp. 628–644). New York, NY: Guilford Press.

D'Mello, S. K., & Mills, C. (in review). *Emotions during writing*.

Damasio, A. (1995). *Descartes' error: Emotion, reason and the human brain*. New York, NY: Quill.

Duran, N. D., Hall, C., McCarthy, P. M., & McNamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, *31*, 439–462. doi:10.1017/S0142716410000068

Isen, A. (2010). Some ways in which positive affect influences decision making. In M. Lewis, J. M Haviland-Jones, & Lisa Feldman-Barrett (Eds.), *Handbook of emotions*, (pp. 548-573). New York, NY: Guilford University Press.

Izard, C. E. (2009). Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual Review of Psychology*, *60*, 1–25. doi:10.1146/annurev.psych.60.110707.163539

Meyer, D., & Turner, J. (2006). Reconceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*, *18*(4), 377–390. doi:10.1007/s10648-006-9032-1

Scherer, K. (2005). What are emotions? And how can they be measured? *Social Sciences Information. Information Sur les Sciences Sociales*, *44*(4), 695–729. doi:10.1177/0539018405058216

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking . In Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.), *Appraisal processes of emotion: Theory, methods, research* (pp. 92–120). New York, NY: Oxford University Press.

Wilson, T., Weibe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. *Proceedings of AAAI*, (pp. 761–769).