Chapter 1
A Joint Introduction to Natural Language Processing and to Deep Learning

Li Deng and Yang Liu

Abstract In this chapter, we set up the fundamental framework for the book. We first provide an introduction to the basics of natural language processing (NLP) as an integral part of artificial intelligence. We then survey the historical development of NLP, spanning over five decades, in terms of three waves. The first two waves arose as rationalism and empiricism, paving ways to the current deep learning wave. The key pillars underlying the deep learning revolution for NLP consist of (1) distributed representations of linguistic entities via embedding, (2) semantic generalization due to the embedding, (3) long-span deep sequence modeling of natural language, (4) hierarchical networks effective for representing linguistic levels from low to high, and (5) end-to-end deep learning methods to jointly solve many NLP tasks. After the survey, several key limitations of current deep learning technology for NLP are analyzed. This analysis leads to five research directions for future advances in NLP.

1.1 Natural Language Processing: The Basics

Natural language processing (NLP) investigates the use of computers to process or to understand human (i.e., natural) languages for the purpose of performing useful tasks. NLP is an interdisciplinary field that combines computational linguistics, computing science, cognitive science, and artificial intelligence. From a scientific perspective, NLP aims to model the cognitive mechanisms underlying the understanding and production of human languages. From an engineering perspective, NLP is concerned with how to develop novel practical applications to facilitate the interactions between computers and human languages. Typical applications in NLP include speech recognition, spoken language understanding, dialogue systems, lexical analysis, parsing, machine translation, knowledge graph, information retrieval, question answering,

L. Deng (✉)
Citadel, Seattle & Chicago, USA
e-mail: l.deng@ieee.org

Y. Liu
Tsinghua University, Beijing, China
e-mail: liuyang2011@tsinghua.edu.cn

L. Deng and Y. Liu (eds.), *Deep Learning in Natural Language Processing*, https://doi.org/10.1007/978-981-10-5209-5_1

sentiment analysis, social computing, natural language generation, and natural language summarization. These NLP application areas form the core content of this book.

Natural language is a system constructed specifically to convey meaning or semantics, and is by its fundamental nature a symbolic or discrete system. The surface or observable "physical" signal of natural language is called text, always in a symbolic form. The text "signal" has its counterpart—the speech signal; the latter can be regarded as the continuous correspondence of symbolic text, both entailing the same latent linguistic hierarchy of natural language. From NLP and signal processing perspectives, speech can be treated as "noisy" versions of text, imposing additional difficulties in its need of "de-noising" when performing the task of understanding the common underlying semantics. Chapters 2 and 3 as well as current Chap. 1 of this book cover the speech aspect of NLP in detail, while the remaining chapters start directly from text in discussing a wide variety of text-oriented tasks that exemplify the pervasive NLP applications enabled by machine learning techniques, notably deep learning.

The symbolic nature of natural language is in stark contrast to the continuous nature of language's neural substrate in the human brain. We will defer this discussion to Sect. 1.6 of this chapter when discussing future challenges of deep learning in NLP. A related contrast is how the symbols of natural language are encoded in several continuous-valued modalities, such as gesture (as in sign language), handwriting (as an image), and, of course, speech. On the one hand, the word as a symbol is used as a "signifier" to refer to a concept or a thing in real world as a "signified" object, necessarily a categorical entity. On the other hand, the continuous modalities that encode symbols of words constitute the external signals sensed by the human perceptual system and transmitted to the brain, which in turn operates in a continuous fashion. While of great theoretical interest, the subject of contrasting the symbolic nature of language versus its continuous rendering and encoding goes beyond the scope of this book.

In the next few sections, we outline and discuss, from a historical perspective, the development of general methodology used to study NLP as a rich interdisciplinary field. Much like several closely related sub- and super-fields such as conversational systems, speech recognition, and artificial intelligence, the development of NLP can be described in terms of three major waves (Deng 2017; Pereira 2017), each of which is elaborated in a separate section next.

1.2 The First Wave: Rationalism

NLP research in its first wave lasted for a long time, dating back to 1950s. In 1950, Alan Turing proposed the Turing test to evaluate a computer's ability to exhibit intelligent behavior indistinguishable from that of a human (Turing 1950). This test is based on natural language conversations between a human and a computer designed to generate human-like responses. In 1954, the Georgetown-IBM experiment demonstrated

the first machine translation system capable of translating more than 60 Russian sentences into English.

The approaches, based on the belief that knowledge of language in the human mind is fixed in advance by generic inheritance, dominated most of NLP research between about 1960 and late 1980s. These approaches have been called rationalist ones (Church 2007). The dominance of rationalist approaches in NLP was mainly due to the widespread acceptance of arguments of Noam Chomsky for an innate language structure and his criticism of N-grams (Chomsky 1957). Postulating that key parts of language are hardwired in the brain at birth as a part of the human genetic inheritance, rationalist approaches endeavored to design hand-crafted rules to incorporate knowledge and reasoning mechanisms into intelligent NLP systems. Up until 1980s, most notably successful NLP systems, such as ELIZA for simulating a Rogerian psychotherapist and MARGIE for structuring real-world information into concept ontologies, were based on complex sets of handwritten rules.

This period coincided approximately with the early development of artificial intelligence, characterized by expert knowledge engineering, where domain experts devised computer programs according to the knowledge about the (very narrow) application domains they have (Nilsson 1982; Winston 1993). The experts designed these programs using symbolic logical rules based on careful representations and engineering of such knowledge. These knowledge-based artificial intelligence systems tend to be effective in solving narrow-domain problems by examining the "head" or most important parameters and reaching a solution about the appropriate action to take in each specific situation. These "head" parameters are identified in advance by human experts, leaving the "tail" parameters and cases untouched. Since they lack learning capability, they have difficulty in generalizing the solutions to new situations and domains. The typical approach during this period is exemplified by the expert system, a computer system that emulates the decision-making ability of a human expert. Such systems are designed to solve complex problems by reasoning about knowledge (Nilsson 1982). The first expert system was created in 1970s and then proliferated in 1980s. The main "algorithm" used was the inference rules in the form of "if-then-else" (Jackson 1998). The main strength of these first-generation artificial intelligence systems is its transparency and interpretability in their (limited) capability in performing logical reasoning. Like NLP systems such as ELIZA and MARGIE, the general expert systems in the early days used hand-crafted expert knowledge which was often effective in narrowly defined problems, although the reasoning could not handle uncertainty that is ubiquitous in practical applications.

In specific NLP application areas of dialogue systems and spoken language understanding, to be described in more detail in Chaps. 2 and 3 of this book, such rationalistic approaches were represented by the pervasive use of symbolic rules and templates (Seneff et al. 1991). The designs were centered on grammatical and ontological constructs, which, while interpretable and easy to debug and update, had experienced severe difficulties in practical deployment. When such systems worked, they often worked beautifully; but unfortunately this happened just not very often and the domains were necessarily limited.

Likewise, speech recognition research and system design, another long-standing NLP and artificial intelligence challenge, during this rationalist era were based heavily on the paradigm of expert knowledge engineering, as elegantly analyzed in (Church and Mercer 1993). During 1970s and early 1980s, the expert system approach to speech recognition was quite popular (Reddy 1976; Zue 1985). However, the lack of abilities to learn from data and to handle uncertainty in reasoning was acutely recognized by researchers, leading to the second wave of speech recognition, NLP, and artificial intelligence described next.

1.3 The Second Wave: Empiricism

The second wave of NLP was characterized by the exploitation of data corpora and of (shallow) machine learning, statistical or otherwise, to make use of such data (Manning and Schtze 1999). As much of the structure of and theory about natural language were discounted or discarded in favor of data-driven methods, the main approaches developed during this era have been called empirical or pragmatic ones (Church and Mercer 1993; Church 2014). With the increasing availability of machine-readable data and steady increase of computational power, empirical approaches have dominated NLP since around 1990. One of the major NLP conferences was even named "Empirical Methods in Natural Language Processing (EMNLP)" to reflect most directly the strongly positive sentiment of NLP researchers during that era toward empirical approaches.

In contrast to rationalist approaches, empirical approaches assume that the human mind only begins with general operations for association, pattern recognition, and generalization. Rich sensory input is required to enable the mind to learn the detailed structure of natural language. Prevalent in linguistics between 1920 and 1960, empiricism has been undergoing a resurgence since 1990. Early empirical approaches to NLP focused on developing generative models such as the hidden Markov model (HMM) (Baum and Petrie 1966), the IBM translation models (Brown et al. 1993), and the head-driven parsing models (Collins 1997) to discover the regularities of languages from large corpora. Since late 1990s, discriminative models have become the *de facto* approach in a variety of NLP tasks. Representative discriminative models and methods in NLP include the maximum entropy model (Ratnaparkhi 1997), supporting vector machines (Vapnik 1998), conditional random fields (Lafferty et al. 2001), maximum mutual information and minimum classification error (He et al. 2008), and perceptron (Collins 2002).

Again, this era of empiricism in NLP was paralleled with corresponding approaches in artificial intelligence as well as in speech recognition and computer vision. It came about after clear evidence that learning and perception capabilities are crucial for complex artificial intelligence systems but missing in the expert systems popular in the previous wave. For example, when DARPA opened its first Grand Challenge for autonomous driving, most vehicles then relied on the knowledge-based artificial intelligence paradigm. Much like speech recognition and NLP, the autonomous driving and

computer vision researchers immediately realized the limitation of the knowledge-based paradigm due to the necessity for machine learning with uncertainty handling and generalization capabilities.

The empiricism in NLP and speech recognition in this second wave was based on data-intensive machine learning, which we now call "shallow" due to the general lack of abstractions constructed by many-layer or "deep" representations of data which would come in the third wave to be described in the next section. In machine learning, researchers do not need to concern with constructing precise and exact rules as required for the knowledge-based NLP and speech systems during the first wave. Rather, they focus on statistical models (Bishop 2006; Murphy 2012) or simple neural networks (Bishop 1995) as an underlying engine. They then automatically learn or "tune" the parameters of the engine using ample training data to make them handle uncertainty, and to attempt to generalize from one condition to another and from one domain to another. The key algorithms and methods for machine learning include EM (expectation-maximization), Bayesian networks, support vector machines, decision trees, and, for neural networks, backpropagation algorithm.

Generally speaking, the machine learning based NLP, speech, and other artificial intelligence systems perform much better than the earlier, knowledge-based counterparts. Successful examples include almost all artificial intelligence tasks in machine perception—speech recognition (Jelinek 1998), face recognition (Viola and Jones 2004), visual object recognition (Fei-Fei and Perona 2005), handwriting recognition (Plamondon and Srihari 2000), and machine translation (Och 2003).

More specifically, in a core NLP application area of machine translation, as to be described in detail in Chap. 6 of this book as well as in (Church and Mercer 1993), the field has switched rather abruptly around 1990 from rationalistic methods outlined in Sect. 1.2 to empirical, largely statistical methods. The availability of sentence-level alignments in the bilingual training data made it possible to acquire surface-level translation knowledge not by rules but from data directly, at the expense of discarding or discounting structured information in natural languages. The most representative work during this wave is that empowered by various versions of IBM translation models (Brown et al. 1993). Subsequent developments during this empiricist era of machine translation further significantly improved the quality of translation systems (Och and Ney 2002; Och 2003; Chiang 2007; He and Deng 2012), but not at the level of massive deployment in real world (which would come after the next, deep learning wave).

In the dialogue and spoken language understanding areas of NLP, this empiricist era was also marked prominently by data-driven machine learning approaches. These approaches were well suited to meet the requirement for quantitative evaluation and concrete deliverables. They focused on broader but shallow, surface-level coverage of text and domains instead of detailed analyses of highly restricted text and domains. The training data were used not to design rules for language understanding and response action from the dialogue systems but to learn parameters of (shallow) statistical or neural models automatically from data. Such learning helped reduce the cost of hand-crafted complex dialogue manager's design, and helped improve robustness against speech recognition errors in the overall spoken language

understanding and dialogue systems; for a review, see He and Deng (2013). More specifically, for the dialogue policy component of dialogue systems, powerful reinforcement learning based on Markov decision processes had been introduced during this era; for a review, see Young et al. (2013). And for spoken language understanding, the dominant methods moved from rule- or template-based ones during the first wave to generative models like hidden Markov models (HMMs) (Wang et al. 2011) to discriminative models like conditional random fields (Tur and Deng 2011).

Similarly, in speech recognition, over close to 30 years from early 1980 s to around 2010, the field was dominated by the (shallow) machine learning paradigm using the statistical generative model based on the HMM integrated with Gaussian mixture models, along with various versions of its generalization (Baker et al. 2009a, b; Deng and O'Shaughnessy 2003; Rabiner and Juang 1993). Among many versions of the generalized HMMs were statistical and neural-network-based hidden dynamic models (Deng 1998; Bridle et al. 1998; Deng and Yu 2007). The former adopted EM and switching extended Kalman filter algorithms for learning model parameters (Ma and Deng 2004; Lee et al. 2004), and the latter used backpropagation (Picone et al. 1999). Both of them made extensive use of multiple latent layers of representations for the generative process of speech waveforms following the long-standing framework of analysis-by-synthesis in human speech perception. More significantly, inverting this "deep" generative process to its counterpart of an end-to-end discriminative process gave rise to the first industrial success of deep learning (Deng et al. 2010, 2013; Hinton et al. 2012), which formed a driving force of the third wave of speech recognition and NLP that will be elaborated next.

1.4 The Third Wave: Deep Learning

While the NLP systems, including speech recognition, language understanding, and machine translation, developed during the second wave performed a lot better and with higher robustness than those during the first wave, they were far from human-level performance and left much to desire. With a few exceptions, the (shallow) machine learning models for NLP often did not have the capacity sufficiently large to absorb the large amounts of training data. Further, the learning algorithms, methods, and infrastructures were not powerful enough. All this changed several years ago, giving rise to the third wave of NLP, propelled by the new paradigm of deep-structured machine learning or deep learning (Bengio 2009; Deng and Yu 2014; LeCun et al. 2015; Goodfellow et al. 2016).

In traditional machine learning, features are designed by humans and feature engineering is a bottleneck, requiring significant human expertise. Concurrently, the associated shallow models lack the representation power and hence the ability to form levels of decomposable abstractions that would automatically disentangle complex factors in shaping the observed language data. Deep learning breaks away the above difficulties by the use of deep, layered model structure, often in the form of neural networks, and the associated end-to-end learning algorithms. The advances in

deep learning are one major driving force behind the current NLP and more general artificial intelligence inflection point and are responsible for the resurgence of neural networks with a wide range of practical, including business, applications (Parloff 2016).

More specifically, despite the success of (shallow) discriminative models in a number of important NLP tasks developed during the second wave, they suffered from the difficulty of covering all regularities in languages by designing features manually with domain expertise. Besides the incompleteness problem, such shallow models also face the sparsity problem as features usually only occur once in the training data, especially for highly sparse high-order features. Therefore, feature design has become one of the major obstacles in statistical NLP before deep learning comes to rescue. Deep learning brings hope for addressing the human feature engineering problem, with a view called "NLP from scratch" (Collobert et al. 2011), which was in early days of deep learning considered highly unconventional. Such deep learning approaches exploit the powerful neural networks that contain multiple hidden layers to solve general machine learning tasks dispensing with feature engineering. Unlike shallow neural networks and related machine learning models, deep neural networks are capable of learning representations from data using a cascade of multiple layers of nonlinear processing units for feature extraction. As higher level features are derived from lower level features, these levels form a hierarchy of concepts.

Deep learning originated from artificial neural networks, which can be viewed as cascading models of cell types inspired by biological neural systems. With the advent of backpropagation algorithm (Rumelhart et al. 1986), training deep neural networks from scratch attracted intensive attention in 1990s. In these early days, without large amounts of training data and without proper design and learning methods, during neural network training the learning signals vanish exponentially with the number of layers (or more rigorously the depth of credit assignment) when propagated from layer to layer, making it difficult to tune connection weights of deep neural networks, especially the recurrent versions. Hinton et al. (2006) initially overcame this problem by using unsupervised pretraining to first learn generally useful feature detectors. Then, the network is further trained by supervised learning to classify labeled data. As a result, it is possible to learn the distribution of a high-level representation using low-level representations. This seminal work marks the revival of neural networks. A variety of network architectures have since been proposed and developed, including deep belief networks (Hinton et al. 2006), stacked auto-encoders (Vincent et al. 2010), deep Boltzmann machines (Hinton and Salakhutdinov 2012), deep convolutional neural works (Krizhevsky et al. 2012), deep stacking networks (Deng et al. 2012), and deep Q-networks (Mnih et al. 2015). Capable of discovering intricate structures in high-dimensional data, deep learning has since 2010 been successfully applied to real-world tasks in artificial intelligence including notably speech recognition (Yu et al. 2010; Hinton et al. 2012), image classification (Krizhevsky et al. 2012; He et al. 2016), and NLP (all chapters in this book). Detailed analyses and reviews of deep learning have been provided in a set of tutorial survey articles (Deng 2014; LeCun et al. 2015; Juang 2016).

As speech recognition is one of core tasks in NLP, we briefly discuss it here due to its importance as the first industrial NLP application in real world impacted strongly by deep learning. Industrial applications of deep learning to large-scale speech recognition started to take off around 2010. The endeavor was initiated with a collaboration between academia and industry, with the original work presented at the 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. The workshop was motivated by the limitations of deep generative models of speech, and the possibility that the big-compute, big-data era warrants a serious exploration of deep neural networks. It was believed then that pretraining DNNs using generative models of deep belief nets based on the contrastive divergence learning algorithm would overcome the main difficulties of neural nets encountered in the 1990s (Dahl et al. 2011; Mohamed et al. 2009). However, early into this research at Microsoft, it was discovered that without contrastive divergence pretraining, but with the use of large amounts of training data together with the deep neural networks designed with corresponding large, context-dependent output layers and with careful engineering, dramatically lower recognition errors could be obtained than then-state-of-the-art (shallow) machine learning systems (Yu et al. 2010, 2011; Dahl et al. 2012). This finding was quickly verified by several other major speech recognition research groups in North America (Hinton et al. 2012; Deng et al. 2013) and subsequently overseas. Further, the nature of recognition errors produced by the two types of systems was found to be characteristically different, offering technical insights into how to integrate deep learning into the existing highly efficient, run-time speech decoding system deployed by major players in speech recognition industry (Yu and Deng 2015; Abdel-Hamid et al. 2014; Xiong et al. 2016; Saon et al. 2017). Nowadays, backpropagation algorithm applied to deep neural nets of various forms is uniformly used in all current state-of-the-art speech recognition systems (Yu and Deng 2015; Amodei et al. 2016; Saon et al. 2017), and all major commercial speech recognition systems—Microsoft Cortana, Xbox, Skype Translator, Amazon Alexa, Google Assistant, Apple Siri, Baidu and iFlyTek voice search, and more—are all based on deep learning methods.

The striking success of speech recognition in 2010–2011 heralded the arrival of the third wave of NLP and artificial intelligence. Quickly following the success of deep learning in speech recognition, computer vision (Krizhevsky et al. 2012) and machine translation (Bahdanau et al. 2015) were taken over by the similar deep learning paradigm. In particular, while the powerful technique of neural embedding of words was developed in as early as 2011 (Bengio et al. 2001), it is not until more than 10 year later it was shown to be practically useful at a large and practically useful scale (Mikolov et al. 2013) due to the availability of big data and faster computation. In addition, a large number of other real-world NLP applications, such as image captioning (Karpathy and Fei-Fei 2015; Fang et al. 2015; Gan et al. 2017), visual question answering (Fei-Fei and Perona 2016), speech understanding (Mesnil et al. 2013), web search (Huang et al. 2013b), and recommendation systems, have been made successful due to deep learning, in addition to many non-NLP tasks including drug discovery and toxicology, customer relationship management, recommendation systems, gesture recognition, medical informatics, advertisement, medical image

analysis, robotics, self-driving vehicles, board and eSports games (e.g., Atari, Go, Poker, and the latest, DOTA2), and so on. For more details, see https://en.wikipedia.org/wiki/deep_learning.

In more specific text-based NLP application areas, machine translation is perhaps impacted the most by deep learning. Advancing from the shallow statistical machine translation developed during the second wave of NLP, the current best machine translation systems in real-world applications are based on deep neural networks. For example, Google announced the first stage of its move to neural machine translation in September 2016 and Microsoft made a similar announcement 2 months later. Facebook has been working on the conversion to neural machine translation for about a year, and by August 2017 it is at full deployment. Details of the deep learning techniques in these state-of-the-art large-scale machine translation systems will be reviewed in Chap. 6.

In the area of spoken language understanding and dialogue systems, deep learning is also making a huge impact. The current popular techniques maintain and expand the statistical methods developed during second-wave era in several ways. Like the empirical, (shallow) machine learning methods, deep learning is also based on data-intensive methods to reduce the cost of hand-crafted complex understanding and dialogue management, to be robust against speech recognition errors under noise environments and against language understanding errors, and to exploit the power of Markov decision processes and reinforcement learning for designing dialogue policy, e.g., (Gasic et al. 2017; Dhingra et al. 2017). Compared with the earlier methods, deep neural network models and representations are much more powerful and they make end-to-end learning possible. However, deep learning has not yet solved the problems of interpretability and domain scalability associated with earlier empirical techniques. Details of the deep learning techniques popular for current spoken language understanding and dialogue systems as well as their challenges will be reviewed in Chaps. 2 and 3.

Two important recent technological breakthroughs brought about in applying deep learning to NLP problems are sequence-to-sequence learning (Sutskevar et al. 2014) and attention modeling (Bahdanau et al. 2015). The sequence-to-sequence learning introduces a powerful idea of using recurrent nets to carry out both encoding and decoding in an end-to-end manner. While attention modeling was initially developed to overcome the difficulty of encoding a long sequence, subsequent developments significantly extended its power to provide highly flexible alignment of two arbitrary sequences that can be learned together with neural network parameters. The key concepts of sequence-to-sequence learning and of attention mechanism boosted the performance of neural machine translation based on distributed word embedding over the best system based on statistical learning and local representations of words and phrases. Soon after this success, these concepts have also been applied successfully to a number of other NLP-related tasks such as image captioning (Karpathy and Fei-Fei 2015; Devlin et al. 2015), speech recognition (Chorowski et al. 2015), meta-learning for program execution, one-shot learning, syntactic parsing, lip reading, text understanding, summarization, and question answering and more.

Setting aside their huge empirical successes, models of neural-network-based deep learning are often simpler and easier to design than the traditional machine learning models developed in the earlier wave. In many applications, deep learning is performed simultaneously for all parts of the model, from feature extraction all the way to prediction, in an end-to-end manner. Another factor contributing to the simplicity of neural network models is that the same model building blocks (i.e., the different types of layers) are generally used in many different applications. Using the same building blocks for a large variety of tasks makes the adaptation of models used for one task or data to another task or data relatively easy. In addition, software toolkits have been developed to allow faster and more efficient implementation of these models. For these reasons, deep neural networks are nowadays a prominent method of choice for a large variety of machine learning and artificial intelligence tasks over large datasets including, prominently, NLP tasks.

Although deep learning has proven effective in reshaping the processing of speech, images, and videos in a revolutionary way, the effectiveness is less clear-cut in intersecting deep learning with text-based NLP despite its empirical successes in a number of practical NLP tasks. In speech, image, and video processing, deep learning effectively addresses the semantic gap problem by learning high-level concepts from raw perceptual data in a direct manner. However, in NLP, stronger theories and structured models on morphology, syntax, and semantics have been advanced to distill the underlying mechanisms of understanding and generation of natural languages, which have not been as easily compatible with neural networks. Compared with speech, image, and video signals, it seems less straightforward to see that the neural representations learned from textual data can provide equally direct insights onto natural language. Therefore, applying neural networks, especially those having sophisticated hierarchical architectures, to NLP has received increasing attention and has become the most active area in both NLP and deep learning communities with highly visible progresses made in recent years (Deng 2016; Manning and Socher 2017). Surveying the advances and analyzing the future directions in deep learning for NLP form the main motivation for us to write this chapter and to create this book, with the desire for the NLP researchers to accelerate the research further in the current fast pace of the progress.

1.5 Transitions from Now to the Future

Before analyzing the future dictions of NLP with more advanced deep learning, here we first summarize the significance of the transition from the past waves of NLP to the present one. We then discuss some clear limitations and challenges of the present deep learning technology for NLP, to pave a way to examining further development that would overcome these limitations for the next wave of innovations.

1.5.1 From Empiricism to Deep Learning: A Revolution

On the surface, the deep learning rising wave discussed in Sect. 1.4 in this chapter appears to be a simple push of the second, empiricist wave of NLP (Sect. 1.3) into an extreme end with bigger data, larger models, and greater computing power. After all, the fundamental approaches developed during both waves are data-driven and are based on machine learning and computation, and have dispensed with human-centric "rationalistic" rules that are often brittle and costly to acquire in practical NLP applications. However, if we analyze these approaches holistically and at a deeper level, we can identify aspects of conceptual revolution moving from empiricist machine learning to deep learning, and can subsequently analyze the future directions of the field (Sect. 1.6). This revolution, in our opinion, is no less significant than the revolution from the earlier rationalist wave to empiricist one as analyzed at the beginning (Church and Mercer 1993) and at the end of the empiricist era (Charniak 2011).

Empiricist machine learning and linguistic data analysis during the second NLP wave started in early 1990 s by crypto-analysts and computer scientists working on natural language sources that are highly limited in vocabulary and application domains. As we discussed in Sect. 1.3, surface-level text observations, i.e., words and their sequences, are counted using discrete probabilistic models without relying on deep structure in natural language. The basic representations were "one-hot" or localist, where no semantic similarity between words was exploited. With restrictions in domains and associated text content, such structure-free representations and empirical models are often sufficient to cover much of what needs to be covered. That is, the shallow, count-based statistical models can naturally do well in limited and specific NLP tasks. But when the domain and content restrictions are lifted for more realistic NLP applications in real-world, count-based models would necessarily become ineffective, no manner how many tricks of smoothing have been invented in an attempt to mitigate the problem of combinatorial counting sparseness. This is where deep learning for NLP truly shines—distributed representations of words via embedding, semantic generalization due to the embedding, longer span deep sequence modeling, and end-to-end learning methods have all contributed to beating empiricist, count-based methods in a wide range of NLP tasks as discussed in Sect. 1.4.

1.5.2 Limitations of Current Deep Learning Technology

Despite the spectacular successes of deep learning in NLP tasks, most notably in speech recognition/understanding, language modeling, and in machine translation, there remain huge challenges. The current deep learning methods based on neural networks as a black box generally lack interpretability, even further away from explainability, in contrast to the "rationalist" paradigm established during the first

NLP wave where the rules devised by experts were naturally explainable. In practice, however, it is highly desirable to explain the predictions from a seemingly "black-box" model, not only for improving the model but for providing the users of the prediction system with interpretations of the suggested actions to take (Koh and Liang 2017).

In a number of applications, deep learning methods have proved to give recognition accuracy close to or exceeding humans, but they require considerably more training data, power consumption, and computing resources than humans. Also, the accuracy results are statistically impressive but often unreliable on the individual basis. Further, most of the current deep learning models have no reasoning and explaining capabilities, making them vulnerable to disastrous failures or attacks without the ability to foresee and thus to prevent them. Moreover, the current NLP models have not taken into account the need for developing and executing goals and plans for decision-making via ultimate NLP systems. A more specific limitation of current NLP methods based on deep learning is their poor abilities for understanding and reasoning inter-sentential relationships, although huge progresses have been made for interwords and phrases within sentences.

As discussed earlier, the success of deep learning in NLP has largely come from a simple strategy thus far—given an NLP task, apply standard sequence models based on (bidirectional) LSTMs, add attention mechanisms if information required in the task needs to flow from another source, and then train the full models in an end-to-end manner. However, while sequence modeling is naturally appropriate for speech, human understanding of natural language (in text form) requires more complex structure than sequence. That is, current sequence-based deep learning systems for NLP can be further advanced by exploiting modularity, structured memories, and recursive, tree-like representations for sentences and larger text (Manning 2016).

To overcome the challenges outlined above and to achieve the ultimate success of NLP as a core artificial intelligence field, both fundamental and applied research are needed. The next new wave of NLP and artificial intelligence will not come until researchers create new paradigmatic, algorithmic, and computation (including hardware) breakthroughs. Here, we outline several high-level directions toward potential breakthroughs.

1.6 Future Directions of NLP

1.6.1 Neural-Symbolic Integration

A potential breakthrough is in developing advanced deep learning models and methods that are more effective than current methods in building, accessing, and exploiting memories and knowledge, including, in particular, common-sense knowledge. It is not clear how to best integrate the current deep learning methods, centered on distributed representations (of everything), with explicit, easily interpretable, and

localist-represented knowledge about natural language and the world and with related reasoning mechanisms.

One path to this goal is to seamlessly combine neural networks and symbolic language systems. These NLP and artificial intelligence systems will aim to discover by themselves the underlying causes or logical rules that shape their prediction and decision-making processes interpretable to human users in symbolic natural language forms. Recently, very preliminary work in this direction made use of an integrated neural-symbolic representation called tensor-product neural memory cells, capable of decoding back to symbolic forms. This structured neural representation is provably lossless in the coded information after extensive learning within the neural-tensor domain (Palangi et al. 2017; Smolensky et al. 2016; Lee et al. 2016). Extensions of such tensor-product representations, when applied to NLP tasks such as machine reading and question answering, are aimed to learn to process and understand massive natural language documents. After learning, the systems will be able not only to answer questions sensibly but also to truly understand what it reads to the extent that it can convey such understanding to human users in providing clues as to what steps have been taken to reach the answer. These steps may be in the form of logical reasoning expressed in natural language which is thus naturally understood by the human users of this type of machine reading and comprehension systems. In our view, natural language understanding is not just to accurately predict an answer from a question with relevant passages or data graphs as its contextual knowledge in a supervised way after seeing many examples of matched questions–passages–answers. Rather, the desired NLP system equipped with real understanding should resemble human cognitive capabilities. As an example of such capabilities (Nguyen et al. 2017)—after an understanding system is trained well, say, in a question answering task (using supervised learning or otherwise), it should master all essential aspects of the observed text material provided to solve the question answering tasks. What such mastering entails is that the learned system can subsequently perform well on other NLP tasks, e.g., translation, summarization, recommendation, etc., without seeing additional paired data such as raw text data with its summary, or parallel English and Chinese texts, etc.

One way to examine the nature of such powerful neural-symbolic systems is to regard them as ones incorporating the strength of the "rationalist" approaches marked by expert reasoning and structure richness popular during the first wave of NLP discussed in Sect. 1.2. Interestingly, prior to the rising of deep learning (third) wave of NLP, (Church 2007) argued that the pendulum from rationalist to empiricist approaches has swung too far at almost the peak of the second NLP wave, and predicted that the new rationalist wave would arrive. However, rather than swinging back to a renewed rationalist era of NLP, deep learning era arrived in full force in just a short period from the time of writing by Church (2007). Instead of adding the rationalist flavor, deep learning has been pushing empiricism of NLP to its pinnacle with big data and big compute, and with conceptually revolutionary ways of representing a sweeping range of linguistic entities by massive parallelism and distributedness, thus drastically enhancing the generalization capability of new-generation NLP models. Only after the sweeping successes of current deep learning methods for NLP

(Sect. 1.4) and subsequent analyses of a series of their limitations, do researchers look into the next wave of NLP—not swinging back to rationalism while abandoning empiricism but developing more advanced deep learning paradigms that would organically integrate the missing essence of rationalism into the structured neural methods that are aimed to approach human cognitive functions for language.

1.6.2 Structure, Memory, and Knowledge

As discussed earlier in this chapter as well as in the current NLP literature (Manning and Socher 2017), NLP researchers at present still have very primitive deep learning methods for exploiting structure and for building and accessing memories or knowledge. While LSTM (with attention) has been pervasively applied to NLP tasks to beat many NLP benchmarks, LSTM is far from a good memory model for human cognition. In particular, LSTM lacks adequate structure for simulating episodic memory, and one key component of human cognitive ability is to retrieve and re-experience aspects of a past novel event or thought. This ability gives rise to one-shot learning skills and can be crucial in reading comprehension of natural language text or speech understanding, as well as reasoning over events described by natural language. Many recent studies have been devoted to better memory modeling, including external memory architectures with supervised learning (Vinyals et al. 2016; Kaiser et al. 2017) and augmented memory architectures with reinforcement learning (Graves et al. 2016; Oh et al. 2016). However, they have not shown general effectiveness, but have suffered from a number of of limitations including notably scalability (arising from the use of attention which has to access every stored element in the memory). Much work remains in the direction of better modeling of memory and exploitation of knowledge for text understanding and reasoning.

1.6.3 Unsupervised and Generative Deep Learning

Another potential breakthrough in deep learning for NLP is in new algorithms for unsupervised deep learning, which makes use of ideally no direct teaching signals paired with inputs (token by token) to guide the learning. Word embedding discussed in Sect. 1.4 can be viewed as a weak form of unsupervised learning, making use of adjacent words as "cost-free" surrogate teaching signals, but for real-world NLP prediction tasks, such as translation, understanding, summarization, etc., such embedding obtained in an "unsupervised manner" has to be fed into another supervised architecture which requires costly teaching signals. In truly unsupervised learning which requires no expensive teaching signals, new types of objective functions and new optimization algorithms are needed, e.g., the objective function for unsupervised learning should not require explicit target label data aligned with the input data as in cross entropy that is most popular for supervised learning. Development of unsu-

pervised deep learning algorithms has been significantly behind that of supervised and reinforcement deep learning where backpropagation and Q-learning algorithms have been reasonably mature.

The most recent preliminary development in unsupervised learning takes the approach of exploiting sequential output structure and advanced optimization methods to alleviate the need for using labels in training prediction systems (Russell and Stefano 2017; Liu et al. 2017). Future advances in unsupervised learning are promising by exploiting new sources of learning signals including the structure of input data and the mapping relationships from input to output and vice versa. Exploiting the relationship from output to input is closely connected to building conditional generative models. To this end, the recent popular topic in deep learning—generative adversarial networks (Goodfellow et al. 2014)—is a highly promising direction where the long-standing concept of analysis-by-synthesis in pattern recognition and machine learning is likely to return to spotlight in the near future in solving NLP tasks in new ways.

Generative adversarial networks have been formulated as neural nets, with dense connectivity among nodes and with no probabilistic setting. On the other hand, probabilistic and Bayesian reasoning, which often takes computational advantage of sparse connections among "nodes" as random variables, has been one of the principal theoretical pillars to machine learning and has been responsible for many NLP methods developed during the empiricist wave of NLP discussed in Sect. 1.3. What is the right interface between deep learning and probabilistic modeling? Can probabilistic thinking help understand deep learning techniques better and motivate new deep learning methods for NLP tasks? How about the other way around? These issues are widely open for future research.

1.6.4 Multimodal and Multitask Deep Learning

Multimodal and multitask deep learning are related learning paradigms, both concerning the exploitation of latent representations in the deep networks pooled from different modalities (e.g., audio, speech, video, images, text, source codes, etc.) or from multiple cross-domain tasks (e.g., point and structured prediction, ranking, recommendation, time-series forecasting, clustering, etc.). Before the deep learning wave, multimodal and multitask learning had been very difficult to be made effective, due to the lack of intermediate representations that share across modalities or tasks. See a most striking example of this contrast for multitask learning—multilingual speech recognition during the empiricist wave (Lin et al. 2008) and during the deep learning wave (Huang et al. 2013a).

Multimodal information can be exploited as low-cost supervision. For instance, standard speech recognition, image recognition, and text classification methods make use of supervision labels within each of the speech, image, and text modalities separately. This, however, is far from how children learn to recognize speech, image, and to classify text. For example, children often get the distant "supervision" signal for

speech sounds by an adult pointing to an image scene, text, or handwriting that is associated with the speech sounds. Similarly, for children learning image categories, they may exploit speech sounds or text as supervision signals. This type of learning that occurs in children can motivate a learning scheme that leverages multimodal data to improve engineering systems for multimodal deep learning. A similarity measure needs to be defined in the same semantic space, which speech, image, and text are all mapped into, via deep neural networks that may be trained using maximum mutual information across different modalities. The huge potential of this scheme has not been explored and found in the NLP literature.

Similar to multimodal deep learning, multitask deep learning can also benefit from leveraging multiple latent levels of representations across tasks or domains. The recent work on joint many-task learning solves a range of NLP tasks—from morphological to syntactic and to semantic levels, within one single, big deep neural network model (Hashimoto et al. 2017). The model predicts different levels of linguistic outputs at successively deep layers, accomplishing standard NLP tasks of tagging, chunking, syntactic parsing, as well as predictions of semantic relatedness and entailment. The strong results obtained using this single, end-to-end learned model point to the direction to solve more challenging NLP tasks in real world as well as tasks beyond NLP.

1.6.5 Meta-learning

A further future direction for fruitful NLP and artificial intelligence research is the paradigm of learning-to-learn or meta-learning. The goal of meta-learning is to learn how to learn new tasks faster by reusing previous experience, instead of treating each new task in isolation and learning to solve each of them from scratch. That is, with the success of meta-learning, we can train a model on a variety of learning tasks, such that it can solve new learning tasks using only a small number of training samples. In our NLP context, successful meta-learning will enable the design of intelligent NLP systems that improve or automatically discover new learning algorithms (e.g., sophisticated optimization algorithms for unsupervised learning), for solving NLP tasks using small amounts of training data.

The study of meta-learning, as a subfield of machine learning, started over three decades ago (Schmidhuber 1987; Hochreiter et al. 2001), but it was not until recent years when deep learning methods reasonably matured that stronger evidence of the potentially huge impact of meta-learning has become apparent. Initial progresses of meta-learning can be seen in various techniques successfully applied to deep learning, including hyper-parameter optimization (Maclaurin et al. 2015), neural network architecture optimization (Wichrowska et al. 2017), and fast reinforcement learning (Finn et al. 2017). The ultimate success of meta-learning in real world will allow the development of algorithms to solve most NLP and computer science problems to be reformulated as a deep learning problem and to be solved by a uniform infrastructure designed for deep learning today. Meta-learning is a powerful

emerging artificial intelligence and deep learning paradigm, which is a fertile research area expected to impact real-world NLP applications.

1.7 Summary

In this chapter, to set up the fundamental framework for the book, we first provided an introduction to the basics of natural language processing (NLP), which is more application-oriented than computational linguistics, both belonging to a field of artificial intelligence and computer science. We survey the historical development of the NLP field, spanning over several decades, in terms of three waves of NLP—starting from rationalism and empiricism, to the current deep learning wave. The goal of the survey is to distill insights from the historical developments that serve to guide future directions.

The conclusion from our three-wave analysis is that the current deep learning technology for NLP is a conceptual and paradigmatic revolution from the NLP technologies developed from the previous two waves. The key pillars underlying this revolution consist of distributed representations of linguistic entities (sub-words, words, phrases, sentences, paragraphs, documents, etc.) via embedding, semantic generalization due to the embedding, long-span deep sequence modeling of language, hierarchical networks effective for representing linguistic levels from low to high, and end-to-end deep learning methods to jointly solve many NLP tasks. None of these were possible before the deep learning wave, not only because of the lack of big data and powerful computation in the previous waves but, equally importantly, due to missing the right framework until the deep learning paradigm emerged in recent years.

After we surveyed the prominent successes of selected NLP application areas attributed to deep learning (with a much more comprehensive coverage of the NLP successful areas in the remaining chapters of this book), we pointed out and analyzed several key limitations of current deep learning technology in general, as well as those for NLP more specifically. This investigation led us to five research directions for future advances in NLP—frameworks for neural-symbolic integration, exploration of better memory models, and better use of knowledge, as well as better deep learning paradigms including unsupervised and generative learning, multimodal and multitask learning, and meta-learning.

In conclusion, deep learning has ushered in a world that gives our NLP field a much brighter future than any time in the past. Deep learning not only provides a powerful modeling framework for representing human cognitive abilities of natural language in computer systems but, as importantly, it has already been creating superior practical results in a number of key application areas of NLP. In the remaining chapters of this book, detailed descriptions of NLP techniques developed using the deep learning framework will be provided, and where possible, benchmark results will be presented contrasting deep learning with more traditional techniques developed before the deep learning tidal wave hit the NLP shore just a few years ago. We

hope this comprehensive set of material will serve as a mark along the way where NLP researchers are developing better and more advanced deep learning methods to overcome some or all the current limitations discussed in this chapter, possibly inspired by the research directions we analyzed here as well.

References

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). *Convolutional neural networks for speech recognition*. IEEE/ACM Trans. on Audio, Speech and Language Processing.

Amodei, D., Ng, A., et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of ICML*.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Baker, J., et al. (2009a). Research developments and directions in speech recognition and understanding. *IEEE Signal Processing Magazine*, *26*(4).

Baker, J., et al. (2009b). Updated MINDS report on speech recognition and understanding. *IEEE Signal Processing Magazine*, *26*(4).

Baum, L., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*.

Bengio, Y. (2009). *Learning Deep Architectures for AI*. Delft: NOW Publishers.

Bengio, Y., Ducharme, R., Vincent, P., & d Jauvin, C. (2001). A neural probabilistic language model. *Proceedings of NIPS*.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Berlin: Springer.

Bridle, J., et al. (1998). An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. *Final Report for 1998 Workshop on Language Engineering, Johns Hopkins University CLSP*.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*.

Charniak, E. (2011). The brain as a statistical inference engine—and you can too. *Computational Linguistics*, *37*.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computaitional Linguistics*.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. In *Proceedings of NIPS*.

Church, K. (2007). A pendulum swung too far. *Linguistic Issues in Language Technology*, *2*(4).

Church, K. (2014). The case for empiricism (with and without statistics). In *Proceedings of Frame Semantics in NLP*.

Church, K., & Mercer, R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, *9*(1).

Collins, M. (1997). *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Reserach*, *12*.

Dahl, G., Yu, D., & Deng, L. (2011). Large-vocabulry continuous speech recognition with context-dependent DBN-HMMs. In *Proceedings of ICASSP*.

Dahl, G., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transaction on Audio, Speech, and Language Processing*, *20*.

Deng, L. (1998). A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, *24*(4).

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, *3*.

Deng, L. (2016). Deep learning: From speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*, *5*.

Deng, L. (2017). Artificial intelligence in the rising wave of deep learning—The historical path and future outlook. In *IEEE Signal Processing Magazine, 35*.

Deng, L., & O'Shaughnessy, D. (2003). *SPEECH PROCESSING A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker.

Deng, L., & Yu, D. (2007). Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition. In *Proceedings of ICASSP*.

Deng, L., & Yu, D. (2014). *Deep Learning: Methods and Applications*. Delft: NOW Publishers.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *Proceedings of ICASSP*.

Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., & Hinton, G. (2010). Binary coding of speech spectrograms using a deep autoencoder. In *Proceedings of Interspeech.*

Deng, L., Yu, D., & Platt, J. (2012). Scalable stacking and learning for building deep architectures. In *Proceedings of ICASSP*.

Devlin, J., et al. (2015). Language models for image captioning: The quirks and what works. In *Proceedings of CVPR.*

Dhingra, B., Li, L., Li, X., Gao, J., Chen, Y., Ahmed, F., & Deng, L. (2017). Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of ACL.*

Fang, H., et al. (2015). From captions to visual concepts and back. In *Proceedings of CVPR.*

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of CVPR.*

Fei-Fei, L., & Perona, P. (2016). Stacked attention networks for image question answering. In *Proceedings of CVPR.*

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML.*

Gan, Z., et al. (2017). Semantic compositional networks for visual captioning. In *Proceedings of CVPR*.

Gasic, M., Mrk, N., Rojas-Barahona, L., Su, P., Ultes, S., Vandyke, D., Wen, T., & Young, S. (2017). Dialogue manager domain adaptation using gaussian process reinforcement learning. *Computer Speech and Language*, *45*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Goodfellow, I., et al. (2014). Generative adversarial networks. In *Proceedings of NIPS.*

Graves, A., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538.*

Hashimoto, K., Xiong, C., Tsuruoka, Y., & Socher, R. (2017). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Proceedings of EMNLP*.

He, X., & Deng, L. (2012). Maximum expected BLEU training of phrase and lexicon translation models. In *Proceedings of ACL.*

He, X., & Deng, L. (2013). Speech-centric information processing: An optimization-oriented approach. *Proceedings of the IEEE*, *101*.

He, X., Deng, L., & Chou, W. (2008). Discriminative learning in sequential pattern recognition. *IEEE Signal Processing Magazine*, *25*(5).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR.*

Hinton, G., & Salakhutdinov, R. (2012). A better way to pre-train deep Boltzmann machines. In *Proceedings of NIPS*.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., & Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*.

Hinton, G., Osindero, S., & Teh, Y. -W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*.

Hochreiter, S., et al. (2001). Learning to learn using gradient descent. In *Proceedings of International Conference on Artificial Neural Networks*.

Huang, P., et al. (2013b). Learning deep structured semantic models for web search using click-through data. *Proceedings of CIKM*.

Huang, J. -T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013a). Cross-lingual knowledge transfer using multilingual deep neural networks with shared hidden layers. In *Proceedings of ICASSP*.

Jackson, P. (1998). *Introduction to Expert Systems*. Boston: Addison-Wesley.

Jelinek, F. (1998). *Statistical Models for Speech Recognition*. Cambridge: MIT Press.

Juang, F. (2016). Deep neural networks a developmental perspective. *APSIPA Transactions on Signal and Information Processing*, *5*.

Kaiser, L., Nachum, O., Roy, A., & Bengio, S. (2017). Learning to remember rare events. In *Proceedings of ICLR*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR*.

Koh, P., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of ICML*.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*.

Lee, L., Attias, H., Deng, L., & Fieguth, P. (2004). A multimodal variational approach to learning and inference in switching state space models. In *Proceedings of ICASSP*.

Lee, M., et al. (2016). Reasoning in vector space: An exploratory study of question answering. In *Proceedings of ICLR*.

Lin, H., Deng, L., Droppo, J., Yu, D., & Acero, A. (2008). Learning methods in multilingual speech recognition. In *NIPS Workshop*.

Liu, Y., Chen, J., & Deng, L. (2017). An unsupervised learning method exploiting sequential output statistics. In arXiv:1702.07817.

Ma, J., & Deng, L. (2004). Target-directed mixture dynamic models for spontaneous speech recognition. *IEEE Transaction on Speech and Audio Processing*, *12*(4).

Maclaurin, D., Duvenaud, D., & Adams, R. (2015). Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of ICML*.

Manning, C. (2016). Computational linguistics and deep learning. In *Computational Linguistics*.

Manning, C., & Schtze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Manning, C., & Socher, R. (2017). *Lectures 17 and 18: Issues and Possible Architectures for NLP; Tackling the Limits of Deep Learning for NLP*. CS224N Course: NLP with Deep Learning.

Mesnil, G., He, X., Deng, L., & Bengio, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Proceedings of Interspeech*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*.

Mohamed, A., Dahl, G., & Hinton, G. (2009). Acoustic modeling using deep belief networks. In *NIPS Workshop on Speech Recognition*.
Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press.
Nguyen, T., et al. (2017). MS MARCO: A human generated machine reading comprehension dataset. arXiv:1611,09268
Nilsson, N. (1982). *Principles of Artificial Intelligence*. Berlin: Springer.
Och, F. (2003). Maximum error rate training in statistical machine translation. In *Proceedings of ACL*.
Och, F., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*.
Oh, J., Chockalingam, V., Singh, S., & Lee, H. (2016). Control of memory, active perception, and action in minecraft. In *Proceedings of ICML*.
Palangi, H., Smolensky, P., He, X., & Deng, L. (2017). Deep learning of grammatically-interpretable representations through question-answering. arXiv:1705.08432
Parloff, R. (2016). Why deep learning is suddenly changing your life. In *Fortune Magazine*.
Pereira, F. (2017). A (computational) linguistic farce in three acts. In http://www.earningmyturns.org.
Picone, J., et al. (1999). Initial evaluation of hidden dynamic models on conversational speech. In *Proceedings of ICASSP*.
Plamondon, R., & Srihari, S. (2000). Online and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*.
Rabiner, L., & Juang, B. -H. (1993). *Fundamentals of Speech Recognition*. USA: Prentice-Hall.
Ratnaparkhi, A. (1997). A simple introduction to maximum entropy models for natural language processing. Technical report, University of Pennsylvania.
Reddy, R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, *64*(4).
Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323*.
Russell, S., & Stefano, E. (2017). Label-free supervision of neural networks with physics and domain knowledge. In *Proceedings of AAAI*.
Saon, G., et al. (2017). English conversational telephone speech recognition by humans and machines. In *Proceedings of ICASSP*.
Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning*. Diploma Thesis, Institute of Informatik, Technical University Munich.
Seneff, S., et al. (1991). Development and preliminary evaluation of the MIT ATIS system. In *Proceedings of HLT*.
Smolensky, P., et al. (2016). Reasoning with tensor product representations. arXiv:1601,02745
Sutskevar, I., Vinyals, O., & Le, Q. (2014). Sequence to sequence learning with neural networks. In *Proceedings of NIPS*.
Tur, G., & Deng, L. (2011). *Intent Determination and Spoken Utterance Classification; Chapter 4 in book: Spoken Language Understanding*. Hoboken: Wiley.
Turing, A. (1950). Computing machinery and intelligence. *Mind*, *14*.
Vapnik, V. (1998). *Statistical Learning Theory*. Hoboken: Wiley.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. -A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, *11*.
Vinyals, O., et al. (2016). Matching networks for one shot learning. In *Proceedings of NIPS*.
Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*.
Wang, Y. -Y., Deng, L., & Acero, A. (2011). *Semantic Frame Based Spoken Language Understanding; Chapter 3 in book: Spoken Language Understanding*. Hoboken: Wiley.
Wichrowska, O., et al. (2017). Learned optimizers that scale and generalize. In *Proceedings of ICML*.
Winston, P. (1993). *Artificial Intelligence*. Boston: Addison-Wesley.

Xiong, W., et al. (2016). Achieving human parity in conversational speech recognition. In *Proceedings of Interspeech*.

Young, S., Gasic, M., Thomson, B., & Williams, J. (2013). Pomdp-based statistical spoken dialogue systems: A review. *Proceedings of the IEEE*, *101*.

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Berlin: Springer.

Yu, D., Deng, L., & Dahl, G. (2010). Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition. In *NIPS Workshop*.

Yu, D., Deng, L., Seide, F., & Li, G. (2011). Discriminative pre-training of deep nerual networks. In *U.S. Patent No. 9,235,799, granted in 2016, filed in 2011*.

Zue, V. (1985). The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, *73*.