Research Statement – Shafiq Joty

The Internet is a great source of human knowledge, but most of the information is in the form of unstructured text. In Natural Language Processing (NLP), we focus on adding structure to this text to uncover relevant information, and to use it in developing end-user application programs. To this end, my primary research goal is twofold: (*i*) developing NLP tools to automatically understand language phenomena that go beyond the individual clauses or sentences of a text, i.e., the discourse structure of the text; and (*ii*) exploiting these discourse analysis tools effectively in downstream NLP applications including machine translation, summarization, question answering, and sentiment analysis. One methodology emphasized throughout my research is to first identify the inherent semantic structures in a given problem, and then to develop structured machine learning models to exploit such structures effectively. My work has relied on deep learning for better representation of the input text and on probabilistic graphical models for capturing dependencies in the output.

A significant part of my discourse research focuses on a special form of discourse called *asynchronous conversations*, i.e., conversations where participants communicate with each other at different times (e.g., forums, emails, twitter). While the number of applications targeting these conversations is growing, sophisticated NLP tools to analyze these conversations are still not sufficiently accurate to support those applications. Also, tools developed for analyzing monologues (e.g., news articles) are not as effective when applied directly to these conversations because the two forms of discourse are different in many aspects. My PhD thesis focused on building novel computational models for different discourse analysis tasks in monologues and in asynchronous conversations, which I describe in Section 1. After PhD, I started working on the applications of discourse analysis, which are described in Section 2. Apart from discourse and its applications, I have also developed novel machine learning models for various NLP applications, which are highlighted in Section 3.

I am also interested in multidisciplinary research that goes beyond NLP, and have developed predictive models for a number of data mining tasks in social networks and in health science. I have recently embarked on joint research projects involving multiple research groups, where my collaborators and I are investigating multilingual (e.g., Mandarine, Tamil) and multimodal (e.g., image, video) language processing problems. Some of these efforts are described in Section 3. Plans for future research that are directly related to my previous work are noted in the respective sections. I have a number of other ideas for future research directions, which I discuss in Section 4.

1 Discourse Analysis

A well-written text is not merely a sequence of independent and isolated sentences, but instead a sequence of structured and related sentences. It addresses a particular topic, often covering multiple subtopics, and is organized in a coherent way that enables the reader to process the information. In discourse analysis we seek to uncover such underlying structures, which can support many downstream applications including machine translation, summarization and information extraction. In my PhD thesis, I proposed novel computational models for discovering the *rhetorical* structure of texts, and the *topical* and the *dialogue* structures of written asynchronous conversations. My PhD thesis was financially supported by the NSERC Alexander Graham Bell Canada Graduate Scholarship (CGS-D), which is awarded to the top-ranked PhD students across Canada.

1.1 Rhetorical Analysis

Clauses in a sentence and sentences in a text are logically connected — the meaning of one relates to that of the previous and the following ones. This logical relation between clauses is called the *coherence structure* of the text. Different formal theories have been proposed to describe this struc-

ture. Rhetorical Structure Theory (RST) is perhaps the most influential one, which posits a tree-like discourse structure. For example, consider the discourse tree in Figure 1 for the following text:

> *But he added: "Some people use the purchasers' index as a leading indicator, some use it as a coincident indicator. But the thing it's supposed to measure — manufacturing strength — it missed altogether last month."*

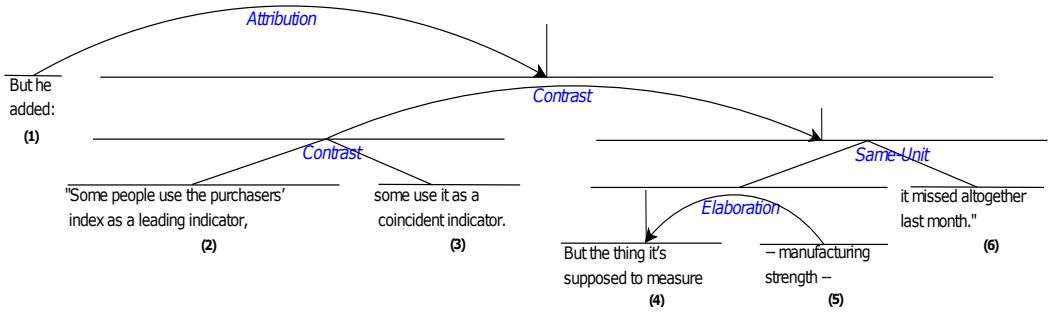

Figure 1: A sample discourse tree. Horizontal lines indicate text segments; satellites are connected to their nuclei by curved arrows and two nuclei are connected with straight lines.

The leaves of a discourse tree correspond to contiguous atomic text spans, called elementary discourse units or EDUs (six in the example). Adjacent EDUs are connected by coherence relations (e.g., *Elaboration*, *Contrast*), forming larger discourse units, which in turn are also subject to this relation linking. Discourse units linked by a relation are further distinguished based on their relative importance in the text: nuclei are the core parts of the relation while satellites are peripheral ones. For example, in Figure 1, the satellite EDU "— manufacturing strength —" *elaborates* the nucleus EDU "But the thing it's supposed to measure", and two nuclei EDUs "Some people use the purchasers index as a leading indicator" and "some use it as a coincident indicator" *contrast* each other. Conventionally, rhetorical analysis involves two subtasks: (*i*) **discourse segmentation** is the task of breaking the text into a sequence of EDUs, and (*ii*) **discourse parsing** is the task of linking the discourse units (EDUs and larger units) into a labeled tree.

Previous approaches to discourse parsing suffer from three key limitations: first, they typically model the structure and the labels of a discourse tree separately in a pipeline fashion, and also do not consider the sequential dependencies between the tree constituents; second, they typically apply greedy and sub-optimal parsing algorithms to build a tree; third, they do not discriminate between intra-sentential parsing (i.e., building the discourse trees for the individual sentences) and multi-sentential parsing (i.e., building the discourse tree for the whole document).

In my PhD dissertation, I developed CODRA – a COmplete Discriminative framework for Rhetorical Analysis [16, 17, 14], which comprises a discourse segmenter and a discourse parser. CODRA addresses the above-mentioned limitations of existing parsers and to the best of my knowledge is still the state-of-the-art and most widely used tool for rhetorical analysis. The crucial component is the use of a probabilistic discriminative parsing model, expressed as a Dynamic Conditional Random Field (DCRF), to infer the probability of all possible tree constituents. By representing the structure and the relation of each tree constituent jointly and by explicitly capturing the sequential dependencies between tree constituents, the DCRF model does not make any independence assumption among these properties. CODRA uses the inferred (posterior) probabilities from the parsing models in a probabilistic CKY-like bottom–up parsing algorithm, which is non-greedy and optimal. Furthermore, a simple modification of this parsing algorithm allows us to generate k-best parse hypotheses, that are later used in a *tree kernel-based reranker* to improve over the initial ranking using additional (global) features of the discourse tree as evidence [23]. I made the source code and a web-based demo of the discourse parser publicly available for research purposes.[1]

Future work: I plan to investigate to what extent discourse segmentation and parsing can be performed jointly. I would also like to explore how CODRA performs on other genres like conversational (e.g., blogs) and evaluative (e.g., reviews) texts. To address the problem of limited annotated

[1] Available at http://alt.qcri.org/tools/discourse-parser/

data in various genres, I am planning to develop an interactive version of CODRA that will allow users to fix the output of the system with minimal effort and let the system learn from that feedback.

1.2 Topic Segmentation and Labeling in Asynchronous Conversations

A discourse, whether it is a monologue or a conversation, exhibits a topic structure. For example, a news article about an earthquake may talk about the intensity, the damage, the aftershocks, and the casualties. Likewise, an email conversation about arranging a conference may discuss conference schedule, organizing committee, accommodation, and registration. **Topic segmentation** refers to the task of grouping the sentences into a set of coherent topical segments, and **topic labeling** is the task of assigning short descriptions to the topical segments to facilitate interpretations of the topics.

While extensive research has been conducted in topic segmentation for monologue and for synchronous dialogue (e.g., meetings), no-one had studied this problem for asynchronous conversations before me. Therefore, there was no reliable annotation scheme, no standard corpus, and no agreed-upon evaluation metrics available. Because of the asynchronous nature, topics in these conversations are often interleaved and do not change in a sequential way as they do in monologue and in synchronous dialogue. As a result, we do not expect models which have proved successful in these domains to be as effective, when directly applied to asynchronous conversations.

In my PhD dissertation, I presented two new corpora of email and blog conversations annotated with topics, and evaluated annotator reliability for the tasks using a new set of metrics, which were also used to evaluate the computational models. I also developed a complete computational framework for performing topic segmentation and labeling in asynchronous conversations [15, 13]. For topic segmentation, I proposed two novel unsupervised models that exploit the fine-grained conversational structure beyond the lexical information. I also proposed a novel graph-theoretic supervised topic segmentation model that combines lexical, conversational and topic features. For topic labeling, I proposed two novel guided random walk models that respectively captures conversation specific clues from two different sources: the leading sentences and the fine-grain conversational structure. Empirical evaluation shows that the segmentation and the labeling performed by the best models outperform the state-of-the-art, and are highly correlated with human annotations. The corpora and the software were made publicly available for research purposes.[2]

Future work: One interesting future direction would be to perform a more extrinsic evaluation of the framework. Instead of testing it with respect to human standards, it would be interesting to see how effective they are in supporting downstream applications, such as conversation summarization and other analytic tasks involving conversations. An example of the later is the ConVis visual analytic system[3] for exploring blog conversations, which uses my framework in the backend.

1.3 Dialogue Act Recognition in Asynchronous Conversation

Apart from the topic and the coherence structures, a conversational discourse (synchronous or asynchronous) also exhibits a dialogue structure; participants interact with each other by performing certain communicative acts like asking questions or requesting something, which are called **dialogue acts**. The two-part structures connecting two acts (e.g., *Question-Answer, Request-Accept*) are called **adjacency pairs**. Previous work on dialogue act modeling has mostly focused on synchronous conversations, and the dominant approaches use supervised sequence taggers like linear-chain CRFs to capture the conversational dependencies (e.g., adjacency pairs) between the act types.

However, modeling conversational dependencies in asynchronous conversation is challenging, because the conversational flow often lacks sequential dependencies in its temporal order. For example, if we arrange the sentences as they arrive in the conversation, it becomes hard to capture any dependency between the act types because the two components of the adjacency pairs can be far apart in the sequence. This leaves us with one open research question: how to model the dependencies between sentences in a single comment and between sentences across different comments?

In my PhD thesis, I proposed *unsupervised* conversational models [12]. First, I showed that like synchronous conversations, it is important for a conversational model in asynchronous conversations to

[2] https://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/Software.html

[3] https://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/ConVis.html

capture the sequential dependencies between the act types. Then, I demonstrated that the conversational models, which are variants of unsupervised Hidden Markov Models (HMMs), learn better sequential dependencies when they are trained on the sequences extracted from the finer conversational structure compared to when they are trained on the temporal order of the sentences. Further investigation shows that the simple unsupervised HMM tends to find topic clusters in addition to dialogue act clusters. To address this problem, I proposed HMM+Mix model which not only explains away the topics, but also improves the act emission distribution by defining it as a mixture model. A part of this work was conducted at Microsoft Research Asia, for which I was given the *"Microsoft Research Excellent Intern"* award.

In a recent work [20], I propose a class of *supervised* structured models in the form of Conditional Random Fields (CRF) defined over arbitrary graph structures of the asynchronous conversation. To surmount the problems with the bag-of-words type representations, my models use sentence representations encoded by a long short term memory (LSTM) recurrent neural model. The LSTM considers the word order and the compositionality of phrases while encoding a sentence vector. Empirical evaluation over three different datasets shows the effectiveness of this approach over existing ones: LSTMs provide better task-specific representations and the global joint model improves over local models. I have also released the source code and the datasets for research purposes.[4]

Future work: I would like to couple CRF with LSTM, so that the LSTM can learn its parameters using the global thread-level feedback. This would require the backpropagation algorithm to take error signals from the global inference algorithm (e.g., loopy belief propagation). We would also like to develop models for conversations, where the conversational structure is given or extractable using the meta data, e.g., reply-to links and usage of quotations in email threads.

1.4 Coherence Modeling

Text analysis models that can distinguish a coherent from incoherent texts are known as coherence models. Such models have a range of applications in text generation, summarization, and coherence scoring. Inspired by the *Centering Theory* [**?**], Barzilay and Lapata [] proposed the popular **entity grid** model of coherence. The model represents a text by a grid (Figure **??**) that captures how grammatical roles of different entities change from sentence to sentence. The grid is then converted into a feature vector containing probabilities of local entity transitions, which enables machine learning models to learn the degree of text coherence. A number of extensions (e.g., incorporating entity-specific features) of this basic entity grid model have been proposed by other researchers.

While the entity grid and its extensions have been successful in many applications, they are limited in several ways. First, they use discrete representation for grammatical roles and other features, which prevents the model from considering sufficiently long transitions. Second, feature vector computation in existing models is decoupled from the target task, which limits the models capacity to learn task-specific features. In our recent work [], we propose a neural architecture for coherence assessment that can capture long range entity transitions along with arbitrary entity- specific features. Our model obtains generalization through distributed representations of entity transitions and entity features. We also present an end-to-end training method to learn task-specific high level features automatically in our model. Our evaluation on three different tasks show that our model achieves state of the art results in all these tasks. We have released our source code for research purposes.

Ongoing work: The above model was proposed for monologic texts (e.g., news articles) and only considers information regarding entities. In our ongoing work, we are extending this model in two different ways. First, we incorporate conversational structure into our model to extend it to asynchronous conversations (e.g., forums).

2 Applications of Discourse Analysis

As mentioned before, discourse analysis has many applications. I am particularly interested in two of them: (*i*) machine translation and (*ii*) text summarization.

[4] http://alt.qcri.org/tools/speech-act/

2.1 Discourse for Machine Translation

Among other applications of discourse, Machine Translation (MT) and its evaluation have received a resurgence of interest recently. Researchers now believe that MT systems should consider discourse phenomena that go beyond the current sentence to ensure consistency in the choice of lexical items or referring expressions, and the fact that source-language coherence relations are also realized in the target language. Automatic MT evaluation is an integral part of the process of developing and tuning MT systems. Reference-based evaluation metrics compare the output of a system to one or more human (reference) translations, and produce a similarity score indicating the quality of the translation. The initial MT metrics approached similarity as a shallow word n-gram matching between the translation and the reference, with a limited use of linguistic information. BLEU is the best-known metric in this family, which has been used for years. However, it has been shown that BLEU and akin metrics are insufficient and unreliable for high-quality translation output.

Modeling discourse brings together the usage of higher-level linguistic information and the exploration of relations beyond the sentence level, which makes it a very attractive goal for MT and its evaluation. In particular, we believe that the semantic and pragmatic information captured in the form of RST trees (e.g., Figure 1) (*i*) can yield better MT evaluation metrics, and (*ii*) can help develop discourse-aware MT systems that produce more coherent translations. In [19, 7], we have explored the first research hypothesis, i.e., (*i*) . Specifically, we show that discourse information can be used to produce evaluation measures that improve over the state-of-the-art in terms of correlation with human assessments. We conduct our research in four steps.

First, we design a simple discourse-aware metric DR-LEX, which use sub-tree kernel to compare RST trees generated with CODRA. We show that this metric helps to improve a large number of MT evaluation measures at the segment-level and at the system-level. Second, we show that tuning the weights in the linear combination of metrics using human assessed examples is a robust way to improve the effectiveness of the DR-LEX metric significantly. Third, we conduct an ablation study which helps us understand which elements of the RST tree have the highest impact on the quality of the evaluation measure. Interestingly enough, the *nuclearity* feature (i.e., the distinction between main and subordinate units) turns out to be more important than the discourse relation labels. Finally, based on these findings, we extend the tree-based representations and present the DISCOTK$_{party}$ metric, which make use of a combination of discourse tree representations and many other metrics. The resulting combined metric with tuned weights scored best as compared to human rankings at the WMT14 Metrics task, both at the system and at the segment levels.

Future work: I would like to explore the potential of discourse information for improving MT systems. My initial plan is to use discourse information to re-rank a set of candidate translations, where the challenge is to establish the links between the discourse structure of the source and that of the translated sentences, trying to promote translations that preserve discourse structure.

2.2 Discourse for Text Summarization

Another important application of discourse structures is text summarization. In my MSc dissertation, I investigated query-focused summarization approaches for answering complex questions (more on this later in Section 3.4). A significant challenge faced by researchers in this field is how to produce summaries that are not only informative but also coherent. I believe discourse structures (e.g., topic, dialogue acts, rhetoric) can play an important role in this aspect. In my ongoing work, I am investigating the utility of discourse structures for conversation (e.g., fora, emails) summarization. The main idea is to impose constraints based on discourse structures in the content selection model to generate summaries that are informative as well as coherent. Discourse trees generated by CODRA allow us to perform content selection at the EDU level. This can yield better compression as compared to existing approaches, which operate at the sentence level.

3 Machine Learning for NLP and Data Mining

Aside from discourse and its applications, I have also worked on developing effective machine learning models for end-user applications. My interests lie in two important sub-fields of machine learning, deep learning and probabilistic graphical models, and their combination.

3.1 Deep Learning

In recent years there has been a growing interest in deep neural networks (DNNs) with application to myriad of NLP and data mining tasks. I have explored DNNs for a number of applications including machine translation and its evaluation, opinion analysis, disaster response, and health informatics.

3.1.1 Deep Learning for Machine Translation and its Evaluation

Machine Translation: A notably successful attempt on using neural networks for Machine Translation (MT) was made by [6]. They proposed a Neural Network Joint Model (NNJM), which augments streams of source with target n-grams and learns a neural model over vector representation of such streams. They achieve impressive gains with NNJM used as an additional feature in the decoder. In [18, 24], we advance the state-of-the-art by extending NNJM for domain adaptation in order to leverage the huge amount of out-of-domain data coming from heterogeneous sources.

We carry out our research in two ways: (*i*) we apply state-of-the-art domain adaptation techniques, such as mixture modeling and data selection using the NNJM, and (*ii*) we propose two novel methods to perform adaptation through instance weighting and weight readjustment in the NNJM framework. Our first method uses data dependent regularization in the loss function to perform (soft) data selection, while the second method fuses the in- and the out-domain models to readjust their parameters. Our evaluation on the standard translation tasks demonstrates that the adapted models outperform the non-adapted baselines and the deep fusion model outperforms the other neural adaptation methods as well as phrase-table adaptation techniques. We also demonstrate that our methods are complementary to the existing methods and together the models can achieve better translation quality. We released our source code in moses open source platform.[5]

Machine Translation Evaluation: In another front [8, 9], we presented a framework for machine translation evaluation using neural networks in a pairwise setting, where the goal is to select the better translation from a pair of hypotheses, given the reference translation. In this framework, lexical, syntactic and semantic information from the reference and the two hypotheses are embedded into small distributed vector representations, and fed into a multi-layer perceptron that models nonlinear interactions between each of the hypotheses and the reference, as well as between the two hypotheses. We experiment with the benchmark datasets from the WMT Metrics shared task, on which we obtain the best results published so far, with the basic network configuration. We also perform a series of experiments to analyze and understand the contribution of the different components of the network. We evaluate variants and extensions including, among others: fine-tuning of the semantic embeddings, and sentence-based representations modeled with recurrent neural networks. The proposed framework is flexible and generalizable, allows for efficient learning and scoring, and provides an MT evaluation metric that correlates with humans on par with the state of the art.

3.1.2 Deep Learning for Opinion Analysis

Fine-grained opinion mining involves: (*i*) identifying the opinion holder, (*ii*) identifying the target or aspect of the opinion, (*iii*) detecting opinion expressions, and (*iv*) measuring the intensity and sentiment of the opinion expressions. For example, in the sentence "John says, the hard disk is very noisy", John, the opinion holder, expresses a very negative opinion towards the target "hard disk" using the opinionated expression "very noisy". In [25], we propose a general class of models based on Recurrent Neural Network (RNN) and word embeddings, that can be successfully applied to fine-grained opinion mining tasks without any task-specific feature engineering effort.

Our results on the task of opinion target extraction show that word embeddings improve the performance of state-of-the-art CRF models, when included as additional features. They also improve RNNs when used as pre-trained word vectors and fine-tuning them on the task gives the best results. A comparison between models demonstrates that RNNs outperform CRFs, even when they use word embeddings as the only features. Incorporating simple linguistic features into RNNs improves the performance even further. Our best results with LSTM RNN outperform the top performing system

[5] http://www.statmt.org/moses/

on the Laptop dataset and achieve the second best on the Restaurant dataset in SemEval-2014. We made our source code publicly available for research.[6]

3.1.3 Deep Learning for Sentence Representation

Vector representation of sentences is important for many text processing tasks that involve clustering, classifying, or ranking sentences. Recently, distributed representation of sentences learned by neural models from unlabeled data has been shown to outperform the traditional bag-of-words representation. However, most of these learning methods consider only the content of a sentence and disregard the relations among sentences by and large. In a recent work [28], we propose a series of novel models for learning latent representations of sentences (i.e., Sen2Vec) that consider the content of a sentence as well as inter-sentence relations. We first represent the inter-sentence relations with a language network and then use the network to induce contextual information into the content-based Sen2Vec models. Two different approaches are introduced to exploit the information in the network. Our first approach *retrofits* (already trained) Sen2Vec vectors with respect to the network in two different ways: (*i*) using the adjacency relations of a node, and (*ii*) using a stochastic sampling method which is more flexible in sampling neighbors of a node. The second approach uses a regularizer to encode the information in the network into the existing Sen2Vec model. Experimental results show that our proposed models outperform existing methods in three fundamental tasks — *classification*, *clustering*, and *ranking*, demonstrating the effectiveness of our approach.

3.1.4 Deep Learning for Health Informatics

Sleep quality is critical for maintaining physical, emotional and mental wellbeing. Although sleep and physical activity are known to be correlated, their relationship is not yet fully understood. The increasing popularity of actigraphy and wearable devices provides a unique opportunity to understand this relationship. In our recent work [29, 30], we explore deep learning models for sleep quality prediction using actigraphy data. In one setting, we first perform human activity recognition (HAR) on raw sensor data, and then feed HAR output into both conventional and deep learning models to perform sleep quality prediction. In the other setting, we employ several deep learning models directly on the raw wearable sensor data without performing HAR or any other feature extraction.

Our results show that using a time-batched LSTM RNN on the raw wearables data improves the sleep quality prediction by an additional 10% with an overall AUC of 0.97 compared to the state-of-the-art non-deep learning approaches, which itself shows a 15% improvement over the current clinical practice. Moreover, utilizing deep learning on raw data eliminates the need for data preprocessing and simplifies the overall workflow to analyze actigraphy data for sleep and physical activity research. From an application impact perspective, the proposed approach promises a very high-fidelity screening test for sleep disorders directly from wearables data, potentially replacing the need for an inconvenient and expensive visit to a sleep laboratory for an evaluation.

3.1.5 Deep Learning for Crisis Computing

During the onset of a crisis situation (e.g., earthquake), rapid analysis of messages posted on microblogging platforms such as Twitter can help humanitarian organizations gain situational awareness, learn about urgent needs, and to direct their decision-making processes accordingly. However, time-critical analysis of such big crisis data brings challenges to machine learning techniques, especially to supervised learning methods. The scarcity of labeled data, particularly in the early hours of a crisis, delays the learning process. Traditional approaches use batch learning with hand engineered features like cue words and TF-IDF vectors. This approach has three major limitations. First, in the beginning of a disaster situation, there is no labeled data available for training for that particular event. Later, the labeled data arrives in minibatches depending on the availability of volunteers. Due to the discrete word representations and the variety across events, traditional classification models perform poorly when trained on previous (out-of-event) events. Second, training a classifier from scratch every time a new minibatch arrives is infeasible. Third, extracting right features for each disaster related classification task is time consuming and requires domain knowledge.

Deep neural networks (DNNs) are ideally suited for disaster response with big crisis data. They are usually trained with online learning and have the flexibility to adaptively learn from new batches of

[6] https://github.com/ppfliu/opinion-target

labeled data without requiring to retrain from scratch. Due to their distributed word representation, they generalize well and make better use of the previously labeled data from other events to speed up the classification process in the beginning of a disaster. DNNs obviate the need of manually crafting features and automatically learn latent features as distributed dense vectors, which generalize well.

In [26], we proposed convolutional neural networks (CNN) for the classification tasks in a disaster situation. CNN captures the most salient n-gram information by means of its convolution and max-pooling operations. On top of the typical CNN, we propose an extension that combines multilayer perceptron with a CNN. We present a series of experiments using different variations of the training data – event data only, out-of-event data only and a concatenation of both. Experiments are conducted for binary (*useful* vs. *not useful*) and multi-class (e.g., *donations*, *sympathy*, *casualties*) classification tasks. Empirical evaluation shows that our CNN models outperform non-neural models by a wide margin in both classification tasks in all scenarios. In the scenario of no event data, the CNN model shows substantial improvement of up to 10 absolute points over several non-neural models. Our variation of the CNN model with multilayer perceptron performed better than its CNN-only counter part. Another finding is that blindly adding out-of-event data either drops the performance or does not give any noticeable improvement over the event only model. To reduce the negative effect of large out-of-event data and to make the most out of it, we apply two simple domain adaptation techniques – (*i*) weight the out-of-event labeled tweets based on their closeness to the event data, (*ii*) select a subset of the out-of-event labeled tweets that are correctly labeled by the event-based classifier. Our results show that the later results in better classification model. We have also released our source code for the crisis computing research community.[7]

3.2 Probabilistic Graphical Models

Probabilistic Graphical Models (PGMs) allow us to define arbitrary joint distributions compactly by exploiting the interactions among the variables, which is necessary for modeling complex dependencies in various NLP and data mining tasks. Apart from the PGMs proposed for discourse analysis tasks in Section 1, I proposed novel PGMs for Community Question Answering (cQA) [11, 21, 22].

In cQA, three tasks are of special relevance when a user poses a new question to the website: (*a*) determine whether a comment within a question-comment thread is a good answer to the question of that thread (i.e., *answer goodness*), (*b*) find related questions to the new question (i.e., *question-question similarity*), and (*c*) find relevant answers to the new question (i.e., *answer selection*). These tasks are interrelated as the information needed to answer a new question is usually found in the good comments of highly related questions.

In [11, 21], I focused on task (*a*) , i.e., classifying comments of an answer-thread as *good* vs. *bad* answers with respect to the thread question. The traditional approach learns a local classifier and uses it to predict for each comment separately. However, this approach ignores the structure in the answer-thread. I approached the task with a global inference process to exploit the information of all comments in the answer-thread in the form of a fully-connected graph. I proposed two novel joint learning models that are on-line and integrate inference within learning. The first one jointly learns two *node*- and *edge*-level MaxEnt classifiers with stochastic gradient descent and integrates the inference step with loopy belief propagation. The second model is an instance of fully connected pairwise CRFs (FCCRF), which performs a global normalization of the functions. The FCCRF model significantly outperforms all other approaches and yields the best results on the task to date. Crucial elements for its success are the global normalization and an Ising-like edge potential.

In a more recent work [22], I consider solving tasks (*b*) and (*c*) jointly with the help of task (*a*) in a joint multi-task learning framework. My approach has two steps. First, a DNN in the form of a feed-forward neural network is trained to solve each of the three individual tasks, and the task-specific hidden layer activations are taken as embedded feature representations to be used in the second step. Then, a structured conditional model, a conditional random field (CRF), uses these embeddings and performs joint learning with global inference to exploit the dependencies between the different tasks.

Our system has been deployed in a real cQA forum site. We have also designed and implemented a web-based interactive cQA interface, which has been evaluated with real forum users (see [10]).

[7] https://github.com/CrisisNLP/deep-learning-for-big-crisis-data

3.3 Combining Deep Learning and Probabilistic Graphical Models

A key strength of deep learning approaches is their ability to learn nonlinear interactions between underlying features through specifically designed hidden layers, and also to learn the features (e.g., word vectors) automatically. In the case of unstructured output problems, this capability has led to gains in many tasks. Deep neural methods are also powerful for structured output problems. Existing work has mostly relied on recurrent or recursive architectures, which can propagate information through hidden layers, but as a result disregard the modeling strength of PGMs, which use global inference to model consistency in the output structure (i.e., class labels of all nodes in a graph).

My current research goal is to combine these two types of models in order to exploit their respective strengths in modeling the input and the output. In my very recent work on speech act recognition (SAR) [20] and on community question answering (cQA) [22], I demonstrate the effectiveness of this marriage of DNNs and PGMs. In both problems, a DNN is first used to encode task-specific embeddings and to perform local classifications which are then "reconciled" in a conditional structured modeling framework, i.e., conditional random fields. However, this integration was done as a two-step process. To make the model more effective, I would like to couple PGMs with DNNs, so that the DNNs can be optimized directly on the global (structured) output.

3.4 Early Work on Unsupervised Models for Question Answering & Summarization

In my M.Sc. thesis, I built an open-domain question answering (QA) system, where I investigated unsupervised methods to automatically answer both simple and complex questions. Simple questions (e.g., "Who is the president of USA?") require small snippets of text as answers and are easier to answer than complex questions (e.g., "Describe the after-effects of cyclone Cindy?") which entail richer information needs and require synthesizing information from multiple documents.

My work on answering complex questions was published in a JAIR article [5] and in one conference paper [4]. I approached the task as a query-focused multi-document summarization and employed an extractive approach to select a subset of the original sentences as the answer. In particular, I experimented with one simple vector space model and two statistical unsupervised models for computing the importance of the sentences. The performance of these approaches depends on the features used and the weighting of these features. I extracted different kinds of informative features for each sentence, and use a gradient descent search to learn the feature-weights from a development set. I first showed that the tree kernel features based on the syntactic and shallow semantic trees of the sentences improve the performance of these models significantly, then I showed that with a large feature set and the optimal feature-weights, my unsupervised models perform as good as state-of-the-art systems with the advantage of not requiring any human annotated data for training. In a separate but related work [3], I show that the syntactic and shallow semantic tree kernels can also improve the performance of the random walk model for answering complex questions.

My approach to answering simple questions was based on question classification and document tagging. Question classification extracts information about how to answer the question (i.e., answer types), and document tagging extracts useful information from the source documents, which are used in finding the answer. To classify the questions based on their answer types (e.g., person, location), I use rules developed manually by analyzing a large set of questions. I employ different off-the-shelf taggers to identify useful information in the documents. In the TREQ-07 QA evaluation, my system was ranked 4th and 6th among 51 participants in the factoid and list type questions, respectively.

4 Future Research Directions

Future directions that are directly related to my previous work are described throughout Sections 1, 2, and 3. In this section, I highlight some of the other ideas that I aspire to explore.

4.1 Other Discourse Analysis Tasks in Asynchronous Conversation

In my discourse-related research, so far I have explored discourse parsing, topic modeling, and dialogue act recognition (Section 1). While discourse parsing can be applied to any form of discourse, the other two address asynchronous conversations. There are other discourse analysis tasks in asyn-

chronous conversations that did not get any attention yet. Two tasks that I am particularly interested in are: (*i*) coreference resolution, and (*ii*) coherence modeling. Although many researchers have studied these tasks in monologue, none has yet addressed them in asynchronous domains. As mentioned before, asynchronous conversation poses a new set of challenges for computational models, and methods designed for monologues often do not work well when applied directly to asynchronous conversation. In my future work I would like to investigate these two tasks for asynchronous conversation by exploiting their conversational structure in joint structured models.

4.2 NLP Applications

I would like to continue exploring new machine learning methods for NLP applications, in particular neural methods for machine translation and summarization. Thus far, my approach to summarization has been *extractive*, where informative sentences are selected to form an abridged version of the source document(s). This approach has been by far the most popular in the field of summarization, largely because it does not require to generate novel sentences. Another approach which is getting more attentions recently is *abstractive* summarization, where the goal is to generate novel texts. Thus, abstractive approach is expected to take us closer to human-style summarization. Very recently, the end-to-end deep learning framework for machine translation (i.e., neural MT) [2] has been used for generating abstractive summaries [27]. While the framework works well for generating short summaries (e.g., headlines), generating longer summaries is still a big challenge. I plan to investigate neural MT for both machine translation and abstractive summarization.

4.3 Group Mining in Social Networks

Beside NLP and its applications, I would be eager to work on data mining tasks in social networks. One particular problem that I am interested in is understanding how groups or teams in social networks behave — group formation and evolution process — which has many real world applications. Groups are best represented with hypergraphs, where hyperedges are used to represent the groups and their interactions. Our proposal is to come up with a *hyperedge2vec* model for learning vector representations for hyperedges, which can then be used in group-related prediction tasks. On the prediction side, our aim is to come up with a conditional structured model to exploit the interactions between sub-groups. We have recently filed a proposal for NSF grant to study this problem [1].

References

[1] A Computational Exploration of the Social Dynamics of Small Groups: Hypergraphs, Tensors, and Deep Learning. *Targeted to the NSF IIS Core program*, Collaboration with the University of Minnesota, Oct, 2016.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] Yllias Chali and Shafiq Joty. Improving the performance of the random walk model for answering complex questions. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL'08, pages 9–12, Columbus, Ohio, 2008. ACL.

[4] Yllias Chali and Shafiq Joty. Selecting sentences for answering complex questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 304–313, Honolulu, Hawaii, 2008. ACL.

[5] Yllias Chali, Shafiq Joty, and Sadid Hasan. Complex question answering: Unsupervised learning approaches and experiments. *Journal of Artificial Intelligence Research*, 35(1):1–47, 2009.

[6] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL'14, pages 1370–1380, Baltimore, USA, 2014. ACL.

[7] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, USA, June 2014. ACL.

[8] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*, ACL'15, pages 805–814, Beijing, China, 2015. ACL.

[9] Francisco Guzman, Shafiq Joty, Lluis Marquez, and Preslav Nakov. Machine Translation Evaluation with Neural Networks. *Journal of Computer Speech and Language (accepted)*, 2017.

[10] Enamul Hoque, Shafiq Joty, Llus Mrquez, and Giuseppe Carenini. CQAVis: Visual Text Analytics for Community Question Answering. In *Proceedings of the 2017 international conference on Intelligent user interfaces*, IUI '17, page to appear, Limassol, Cyprus, 2017. ACM.

[11] Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Global thread-level inference for comment classification in community question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 573–578, Lisbon, Portugal, 2015. ACL.

[12] Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the twenty second International Joint Conference on Artificial Intelligence*, IJCAI'11, Barcelona, Spain, 2011.

[13] Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng. Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, EMNLP'10, pages 388–398, Massachusetts, USA, 2010. ACL.

[14] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 904–915, Jeju Island, Korea, 2012. ACL.

[15] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. Topic Segmentation and Labeling in Asynchronous Conversations. *Journal of Artificial Intelligence Research*, 47:521–573, 2013.

[16] Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41:3:385–435, 2015.

[17] Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 486–496, Sofia, Bulgaria, 2013. ACL.

[18] Shafiq Joty, Nadir Durrani, Hassan Sajjad, and Ahmed Abdelali. Domain Adaptation Using Neural Network Joint Model. *Journal of Computer Speech and Language (accepted)*.

[19] Shafiq Joty, Francisco Guzman, Lluis Marquez, and Preslav Nakov. Discourse Structure in Machine Translation Evaluation. *Computational Linguistics*, xx:x:xx–xx (under review), 2016.

[20] Shafiq Joty and Enamul Hoque. Speech Act Modeling of Written Asynchronous Conversations with Task-Specific Embeddings and Conditional Structured Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 1746–1756. ACL, 2016.

[21] Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Joint Learning with Global Inference for Comment Classification in Community Question Answering. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'16, San Diego, California, 2016.

[22] Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Joint Multitask Learning for Community Question Answering Using Task-Specific Embeddings. *Transactions of the Association for Computational Linguistics*, xx:x:xx–xx (under review), 2016.

[23] Shafiq Joty and Alessandro Moschitti. Discriminative reranking of discourse parses using tree kernels. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2049–2060, Doha, Qatar, October 2014. ACL.

[24] Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, pages 1259–1270, Lisbon, Portugal, 2015. ACL.

[25] Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, pages 1433–1443, Lisbon, Portugal, 2015.

[26] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks. In *arxiv.org/pdf/1608.03902v1.pdf*, (Submitted), 2016.

[27] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[28] Tanay Kumar Saha, Shafiq Joty, Naeemul Hassan, and Mohammad Al Hasan. Dis-S2V: Discourse Informed Sen2Vec. In *Proceedings of the 26th international conference on World Wide Web (under review)*, WWW'17, pages x–x. ACM, 2017.

[29] Aarti Sathyanarayana, Shafiq Joty, Luis Fernandez-Luque, Ferda Ofli, Jaideep Srivastava, Ahmed Elmagarmid, Shahrad Taheri, and Teresa Arora. Sleep Quality Prediction From Wearable Data Using Deep Learning. *JMIR mHealth and uHealth (JMU)*, 4(4)(e125), 2016.

[30] Aarti Sathyanarayana, Shafiq Joty, Ferda Ofli, and Jaideep Srivastava. Impact of Physical Activity on Sleep:A Deep Learning Based Exploration. In *SIAM International Conference on Data Mining*, SDM '17, page (submitted), 2017.