Identifying Referential Intention with Heterogeneous Contexts

Wenhao Yu
University of Notre Dame
Notre Dame, IN 46556
wyu1@nd.edu

Mengxia Yu
University of Notre Dame
Notre Dame, IN 46556
myu2@nd.edu

Tong Zhao
University of Notre Dame
Notre Dame, IN 46556
tzhao2@nd.edu

Meng Jiang
University of Notre Dame
Notre Dame, IN 46556
mjiang2@nd.edu

ABSTRACT

Citing, quoting, and forwarding & commenting behaviors are widely seen in academia, news media, and social media. Existing behavior modeling approaches focused on mining content and describing preferences of authors, speakers, and users. However, behavioral intention plays an important role in generating content on the platforms. In this work, we propose to identify the referential intention which motivates the action of using the referred (e.g., cited, quoted, and retweeted) source and content to support their claims. We adopt a theory in sociology to develop a schema of four types of intentions. The challenge lies in the heterogeneity of observed contextual information surrounding the referential behavior, such as referred content (e.g., a cited paper), local context (e.g., the sentence citing the paper), neighboring context (e.g., the former and latter sentences), and network context (e.g., the academic network of authors, affiliations, and keywords). We propose a new neural framework with Interactive Hierarchical Attention (IHA) to identify the intention of referential behavior by properly aggregating the heterogeneous contexts. Experiments demonstrate that the proposed method can effectively identify the type of intention of citing behaviors (on academic data) and retweeting behaviors (on Twitter). And learning the heterogeneous contexts collectively can improve the performance. This work opens a door for understanding content generation from a fundamental perspective of behavior sciences.

KEYWORDS

Referential Intention, Heterogeneous Contexts, Interactive Hierarchical Attention

ACM Reference Format:
Wenhao Yu, Mengxia Yu, Tong Zhao, and Meng Jiang. 2020. Identifying Referential Intention with Heterogeneous Contexts. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380175

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380175

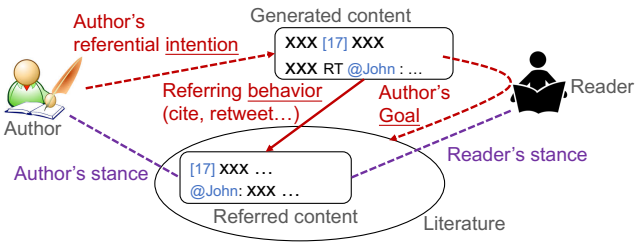


Figure 1: The goal of generating content is to make author's claims/ideas accepted by readers and existing literature. To achieve it, referential behavior such as citing and retweeting uses existing contents in the literature as a bridge between the author and readers. So the author's referential intention depends on author's stance and reader's stance to the referred content. Given observed generated/referred content and a referential behavior, our work aims at identifying the referential intention to understand contents and stances. In this figure, dashed lines are unobservable and to be inferred; only the solid arrow and contents are observable.

1 INTRODUCTION

Content is not isolated from the time of being generated. One of the most important purposes of generating contents is to make the facts, claims, ideas, and/or opinions in the content accepted by readers. For example, researchers write papers to present new ideas, observations, and methods for being used and/or cited; journalists write news articles to broadcast factual information; social media users post messages to express and share their opinions. (We use "authors" to simplify mentions of researchers, journalists, social media users, and many others who generate contents for reading.)

One of the most effective methods to shorten the distance between author and readers is to *refer* to related sources in the generated content (**GC**). We call it *referential behavior*, including citing, quoting, forwarding, and commenting behaviors that can be widely seen in academia, news media, and social media. The referred content (**RC**) acts as a natural bridge, because it has been publicized as literature, news archives, or trending topics. For example, researchers can discuss advantages or major flaws of a cited work to support their claims; users can explain why they like or dislike a shared message to show their opinions on the topic or event. Clearly, the intention of referential behavior depends on author's stance

Referential intentions		Reader's Stance		
		Positive	Neutral	Negative
Author's Stance	Positive	Accept	Strong Acc. (SA)	SA
	Neutral	Background (BG)	BG	BG
	Negative	Strong Rej. (SR)	SR	Reject

Table 1: Given author's different stance and reader's different stance to referred content, the intention of the author's referential behavior, underlying why and how the referred content was used, may be different. According to a theory in sociology [4], we categorize the intention into five: Strong Accept, Accept, Background, Reject, and Strong Reject. The author's generated content, referred content, the referential behavior, and many other contexts reflect the intention.

and readers' stance (in general) to RC, which becomes the most important factor of shaping GC. This observation is supported by a theory in sociology developed in 1980s [4]. Figure 1 presents their relationships. Note that only the referential behaviors including GC and RC are observable. Identifying the underlying referential intention can help understand the process of content generation, infer author's stance and reader's stance, find truth and evidence, and detect misbehavior and misinformation.

In this work, we study author's referential intention and develop a computational method to identify the intention of referential behaviors on multiple platforms such as academia, news media, and social media. Inspired by [4], we build a schema of 5 intention categories: Strong Accept (SA), Accept (A), Background (BG), Reject (R), and Strong Reject (SR). Table 1 presents the relationship between intention category and author/reader's stance to RC.

Author's stance to RC has three types: (1) *Positive* means GC and RC have the same or similar viewpoints; (2) *Neutral* means GC and RC are related but GC does not discuss RC's viewpoint *or* RC has no viewpoint; (3) *Negative* means GC and RC have opposing viewpoints. Take citing behaviors as examples. Suppose GC introduced a new topic model. If RC was a sampling method adopted in the model, the stance would be positive; if RC was the definition of graphical model, it would be neutral; if RC was an existing topic model and GC discussed its weak points, the stance would be negative. Similar can be observed on social media. Suppose GC was on the cons of Obamacare. If RC was about "reasons to still hate Obamacare", the stance would be positive; if RC was generally about healthcare system or insurance policy, it would be neutral; if RC was on the good points, the stance would be negative.

Reader's stance to RC also has three types; however, they are author's assumption on his/her general readers/audience. Though the type names are the same as those of author's stance, their definitions are very different. (1) *Positive* means RC is acceptable by general audience because it has been widely adopted or used in literature *or* it has been widely broadcast on mass media and social media, for instance, survey on deep neural networks *or* fear of financial crisis. (2) *Neutral* means RC has no viewpoint *or* neither RC's viewpoint nor the opposite dominates, for instance, good/bad points of Obamacare. (3) *Negative* means RC is wrong in commonsense, for instance, anything that legitimizes racial discrimination.

Based on the author/reader's stance types, we define intention categories (see Table 1) along with examples of citing behaviors (in an AAAI'16 paper titled "Deep neural networks for learning graph representations"; see Table 2) and retweeting behaviors.

Strong Accept (SA): Table 1 shows that when the author's stance is positive and the reader's stance is non-positive (neutral or negative), the author's intention is to present a strong acceptance of RC to readers. Take the first row of Table 2 as an example. At the time paper [5] (GC) was submitted, *word2vec* (RC) (including the skip-gram model and representation learning idea) has not yet been widely accepted. So, when the authors of [5] used *word2vec* to generate representations, they presented not only the usage but also the advantages of using the method in their work. The authors were trying to make readers accept the advantages (which were not widely accepted) along with the usage. On social media, supportive comments are common in retweets such as listing a presidential candidate's achievements on news about his/her campaign.

Accept (A): Suppose the author's stance is positive and the reader's stance is likely to be positive too, which means RC is widely accepted by large audience for a long time. The author doesn't have to or wouldn't like to sell RC because it cannot be or relate to any interesting point of the GC (e.g., paper's contribution). In the example, because SVD has been commonly used to compress matrix, the authors presented the usage *without* further explanation or comments to it *but* concluded the section. On social media, when the RC is about disaster response to a flood, a fire, or a collapse of buildings in an earthquake, users who retweeted this message would assume that all readers believe the response is necessary.

Background (BG): The author's stance is neutral when GC and RC have related topics but the RC was just used to provide background information but not used to support the claims/ideas in GC. The RC can be a survey paper, a representative work of a topic, or factual news reports, or Wiki records.

Strong Reject (SR) and Reject (R): When the author's stance is negative, GC is to reject RC – SR is to point out the shortcomings of certain methods/techniques, bad outcomes of certain policies, or poor performance of a person or an organization. The rejection will support the author's idea of addressing the problem or adjusting the policy. Interestingly, when it comes to be a rejection, though for some cases all readers are negative to the RC, such as Hitler and World War II, the GC is always a Strong Reject. **So we merge SR and R together as SR.**

Identifying referential intention is non-trivial and challenging. We recruited three people to label an academic dataset (40 papers and 1,565 citations, on two research topics – network embedding and language model) and a general dataset (48 news articles, 249 tweets, and 401 referential links, on 20 events). 1,497 (95.7%) of the citations and 368 (91.8%) of the referential links have consistent labels. So, referential intentions exist and are determinable. However, all the annotators conclude that they *have to* read/know a lot of contexts related to the referential behavior for high confidence of labelling. The work was difficult and time-consuming.

To automatically identify the intention with a computational model, the main challenge is on modeling the heterogeneous contexts surrounding referential behavior. Take citing behavior as an example. The intention is related to *referred content* (e.g., a cited paper), *local context* (LC) of generated content (e.g., the sentence citing the paper), *neighboring contexts* (NC) of GC (e.g., the former and latter sentences), and *network context* (NetC) (e.g., the academic

Generated Content (GC)	Referred Content (RC)	Intention
"Next, the skip-gram model proposed by [22] is used to learn low-dimensional representations for vertices from such linear structures. The learned vertex representations were shown to be effective across a few tasks, outperforming several previous approaches such as..."	[22] Mikolov et al. NeurIPS (2013). (known as word2vec) "Distributed representations of words and phrases and their compositionality."	Strong Accept (SA)
"SVD [17] is a common matrix factorization method that is used to reduce dimensions or extract features. Following [17], we used SVD to compress the PPMI matrix to obtain low dimensional representations. □ (end of section) *5.1 Parameters* ..."	[17] Levy et al. NeurIPS (2014). "Neural word embedding as implicit matrix factorization"	Accept (A)
"Recently, there has been significant interest in the work of learning word embeddings [5]. Their goal is to learn for each natural language word a low-dimensional vector representation based on their contexts, from a large amount of natural language texts."	[5] Bullinaria et al. BRM (2007). "Extracting semantic representations from word co-occurrence statistics: A computational study."	Background (BG)
"An example of a matrix factorization method is hyperspace analogue analysis [20]. A major shortcoming of such an approach and related methods is that frequent words with relatively little semantic value such as stop words have a disproportionate effect on the word representations generated. ... to address this problem ..."	[20] Lund et al. BRM (1996). "Producing high-dimensional semantic spaces from lexical co-occurrence"	Strong Reject (SR)

Table 2: Examples of citing behaviors in each intention category. The authors of GC [5] assumed the readers' stance to the RCs. Expressions based on the author's intent have been highlighted as underlined plus blue colored. "SA" sells RC's advantages assuming that RC has not yet been widely accepted. "A" presents the acceptance on RC but not mentions the reason because it assumes RC is widely accepted. The black underlined are related content (concepts, methods, etc.) between GC and RC.

network of authors, affiliations, and keywords). On one hand, most of the contexts were long (sentences) or big (networks) while only part(s) of them could be strongly related to the intention. On the other hand, we (and the annotators) observe that the multiple types of contexts have interaction with each other: LC and NC come from the same document and form semantic ordering; RC and LC were tightly related on content; RC and LC were produced by the nodes in NetC, and so on. A flat neural framework with the contexts fed but isolated could not perform well.

In this work, we propose a new framework for identifying referential intentions by modeling heterogeneous contexts. Firstly, we use language model to transform text-based contexts (LC, NC, and RC) into embeddings and also use graph neural network to learn author and GC embeddings with NetC. Secondly, we apply local attention mechanism to discover the significant parts corresponding to the intention from each type of text-based context embeddings. Lastly, we design a weighted pooling layer plus a hierarchical attention layer for properly pairing interactive contexts of local attentions into joint global attention.

Experiments on the two real datasets demonstrate that our proposed method of the Interactive Hierarchical Attention (IHA) mechanism outperforms non-attention and non-interactive attention methods: F1 scores were improved relatively by 7.8% on the academic dataset and by 12.8% on the news/social media dataset. Accurate information of referential intentions can enrich the academic and general knowledge graphs (putting a new, important type of attributes on the referential edges between contents) and help infer the truth and find evidence.

The main contributions of this work are summarized as follows.

- We propose a new problem in user/behavior modeling – identifying intention of referential behaviors which are common in academia, news and social media. We define an intention schema based on a theory in behavioral and social sciences.
- We propose a new mechanism called Interactive Hierarchical Attention and a new framework for modeling the relationship between the referential intention and naturally interactive, heterogeneous contexts.
- Experiments on real-world, released datasets of two domains demonstrate the effectiveness of our proposed mechanism and framework. The work will provide a different angle for computational methods to understand content generation.

2 PROBLEM DEFINITION

In this section, we first formally define the research problem and then explain several concepts used in the problem definition.

Problem (Referential intention identification). *Given a referential behavior $b \in \mathcal{B}$, our goal is to learn a mapping function $g(b) : \mathcal{B} \rightarrow \mathcal{S}$, where $\mathcal{B}$ is the set of behaviors and $\mathcal{S}$ is the schema of four types referential intentions: $\mathcal{S}$ = {Strong Accept (SA), Accept (A), Background (BG), Strong Reject (SR)}.*

To develop a learning model for this problem, we need to extract information (and then use the model to turn into features) from the referential behavior. *A referential behavior b has three components which could be represented as $b = \{GC(b), RC(b), NetC(b)\}$:*

Note that the behavior was created by "author" $v_a(b) \in \mathcal{V}_a$ based on the author's intention. $\mathcal{V}_a$ is the set of authors.

(1) Generated content GC(b). The author created the *text* to share their ideas, opinions, claims, or experience with readers. In the text, the author gave contextual information and referred to outside information for readers to understand better and become more likely to accept his/her point of view.

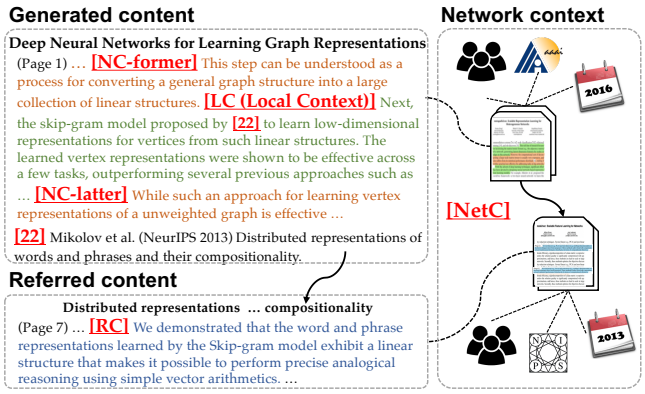


Figure 2: Given a referential behavior in academic articles (i.e., a citation), five types of contexts are associated with the behavior's intention: Generated content includes "Neighboring former context" (NC-former), "Local context" (LC), and "Neighboring latter context" (NC-latter); Referred content (RC); and Network context (NetC) of the academic network of author, venue, and keyword nodes.

- *Local context LC(b):* It has local *text* information surrounding the position of referential behavior (e.g., the brackets "[]" in scientific papers, the hyperlinks in news articles).
- *Neighboring former context NC-former(b):* (simplified as NC-f(b)) It has *text* information before the local context. It is often about former steps of a method or important concepts terms that will be used in the local context.
- *Neighboring latter context NC-latter(b):* (simplified as NC-l(b)) It has *text* information after the local context. It often makes additional comments or explanations to the local context.

(2) Referred content RC(b). It is the *text* the author referred to in the local context of his/her generated content. It is supposed to have strongly-related content with the local context.

(3) Network context NetC(b). It is actually a sub-network of the complete heterogeneous network $G = (\mathcal{V} = \mathcal{V}_a \cup \mathcal{V}_c \cup \mathcal{V}_l \cup \ldots, \mathcal{E})$, where $\mathcal{V}_a$ is the author node set, $\mathcal{V}_c$ is the set of content nodes (e.g., papers, news, tweets), $\mathcal{V}_l$ is the set of "location" nodes (e.g., conferences/journals, geotags), and $\mathcal{E}$ has the links between the nodes. NetC(b) was formed by a sub-network of nodes within 2 hops from author node $v_a(b)$, generated content nodes GC(b), and referred content node RC(b).

Figure 2 presents an example of citing behavior to explain the five types of contexts (LC, NC-former, NC-latter, RC, and NetC). For news articles, hyperlinks of certain terms or phrases can be considered as referential behaviors; we can also find the five types for each of the behaviors. For retweets, we have three types: LC (i.e., the comment added when forwarding), RC (i.e., the forwarded tweet), and NetC (i.e., the social and information network).

Note that any learning model for identifying the intention type has to learn the relationship between the intention and the *heterogeneous* behavior contexts including texts and graphs. The model must be able to extract information from both unstructured and structured data; and it still faces *two challenges*. One is that only part(s) of the texts (in LC or NC or RC) and graphs (in NetC), not all the words or nodes or links, were associated with the referential intention. The other is that the heterogeneous contexts are not isolated but interactive in pairs. For example, LC and NC have semantic ordering; LC and RC are somehow summary and full description of a thing, respectively; and even LC acts as rich attributes of a node in NetC. In next section, we propose a framework with a new mechanism called Interactive Hierarchical Attention to learn $g(\cdot)$ from data of heterogeneous contexts.

3 THE PROPOSED FRAMEWORK

In this section, we first present the overview of our method. Then we introduce its details such as encoders, local attentions, and interactive hierarchical attention for modeling heterogeneous contexts.

Overview: Our proposed method has three parts: first, it encodes heterogeneous contexts using language model for text contexts (NC-f, LC, NC-l, and RC) and using heterogeneous GNN [29] for network context; second, it applies sequence attention [22] for text contexts and graph attention [23] for network context to discover important part(s) associated with intention from each type of input; third, it has a novel mechanism called interactive hierarchical attention that models pairing effects of heterogeneous contexts and merges into global attention. Finally, a softmax function matches the output into four types of referential intentions.

3.1 Encoders for Heterogeneous Contexts

Text encoders. For textual contexts of a referential behavior b, such as LC(b), NC(b), and RC(b), the goal is to learn word embeddings: $\mathbf{h}_{LC(b)_i} \in \mathbb{R}^d$, $\mathbf{h}_{NC(b)_i} \in \mathbb{R}^d$, and $\mathbf{h}_{RC(b)_i} \in \mathbb{R}^d$ for the i-th word in the context, where d is the number of dimensions of embedding vectors. We use $TextC(b)$ to generally denote a type of context. We use BiLSTM to learn contextual word embeddings. It has a forward LSTM which reads text $TextC(b)$ from word $TextC(b)_1$ to $TextC(b)_T$, where T is the length of the text. It also has a backward LSTM to learn from the other direction. Then we have:

$$\mathbf{h}_{TextC(b)_i} = \overrightarrow{LSTM}(TextC(b)_i) \oplus \overleftarrow{LSTM}(TextC(b)_i), i = 1 \ldots T, \quad (1)$$

where operation $\oplus$ denotes concatenation.

Graph encoders. Given heterogeneous network context $NC(b) = (\mathcal{V}, \mathcal{E})$, the goal is to learn: (1) an embedding vector of node's textual attributes (e.g., keyword, venue's name): $\mathbf{v}^{(attribute)} \in \mathbb{R}^d$ for each node $v \in \mathcal{V}$ and (2) an embedding vector of aggregating representations of node v's neighbors of type t (e.g., author, venue, keyword, paper): $\mathbf{v}^{(neighbor),t} \in \mathbb{R}^d$:

$$\mathbf{v}^{(neighbor),t} = \mathcal{AG}^t_{u \in \mathcal{N}_t(v)}\{\mathbf{u}^{(attribute)}\}, \quad (2)$$

where $\mathcal{N}_t(v)$ is the set of node v's neighbors of type t, $\mathcal{AG}$ is an aggregator (e.g., mean pooling, LSTM). We adopt heterogeneous graph neural network [29] as graph encoders to learn the embeddings. These vectors will be used in the next part, graph attention.

3.2 Local Attentions for Heterogeneous Contexts

Sequence attention. Not all words contribute equally to the representation of the sentence meaning. For example, a turning word

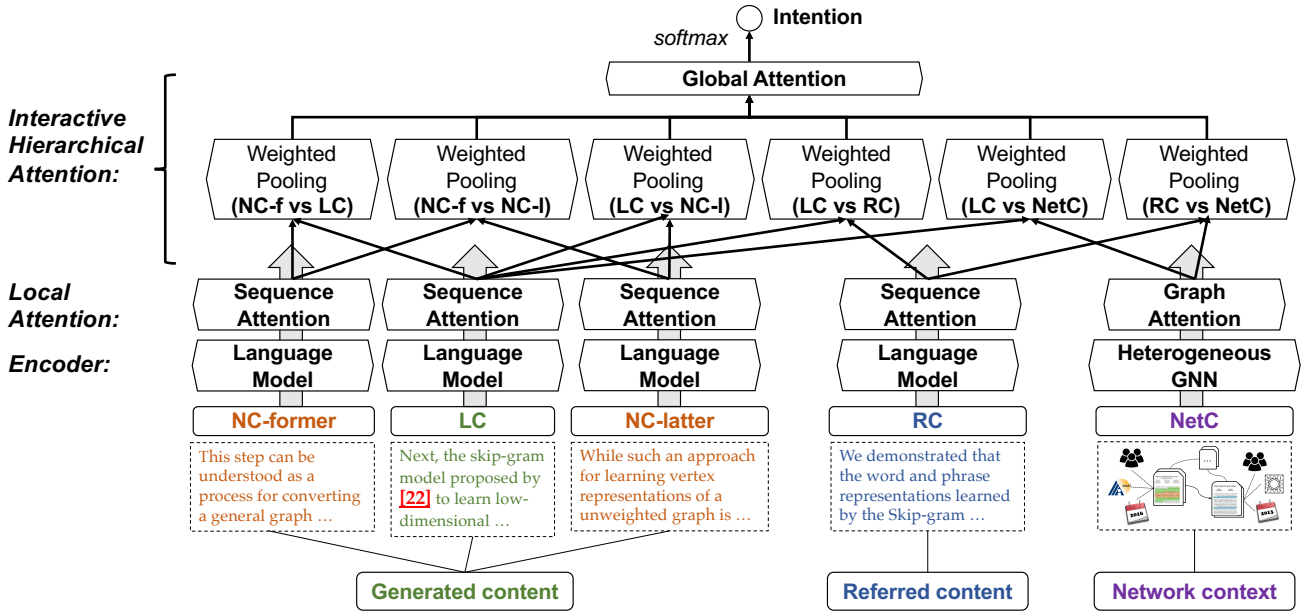


Figure 3: Our proposed framework has three parts: first, it encodes heterogeneous contexts - using language model (e.g., LSTM) for text contexts (NC-f, LC, NC-l, and RC) and using heterogeneous GNN [29] for network context; second, it applies sequence attention for text contexts [22] and graph attention for network context [23] to discover important part(s) associated with intention from each type of input; third, it has a novel mechanism called interactive hierarchical attention that models pairing effects of heterogeneous contexts and merges into global attention.

(e.g., "however", "nevertheless") or a word reflecting the author's stance (e.g., "state-of-the-art", "shortcoming") contributes more significantly than other words. Hence, we use attention mechanisms to learn the importance of such words on the meaning of the sentences and aggregate the representation of those informative words to represent each type of textual context $TextC(b)$ (e.g., LC, NC, RC) with a numerical vector. Formally, we have

$$\mathbf{h}_{TextC(b)} = \sum_i \alpha_{TextC(b)_i} \cdot \mathbf{h}_{TextC(b)_i}, \tag{3}$$

where $\mathbf{h}_{TextC(b)_i}$ is given by Eq.(1) and $\alpha_{TextC(b)_i}$ is the attention weight of the i-th word in textual context $TextC(b)$:

$$\alpha_{TextC(b)_i} = \frac{exp(\beta_{TextC(b)_i})}{\sum_i exp(\beta_{TextC(b)_i})}, \ \beta_{TextC(b)_i} = f(\mathbf{h}_{TextC(b)_i}), \tag{4}$$

where $f(\cdot)$ is a function (e.g., weighted hyperbolic tangent, multilayer perceptron) that transforms the hidden representation $h_{TextC(b)_i}$ to a scalar weight $\beta_{TextC(b)_i}$.

Graph attention. For each node $v \in \mathcal{V}$, the goal is to learn the importance of different embeddings given in Section 3.1 and generate a universal embedding for the node:

$$\mathbf{v} = \alpha^{v,v}\mathbf{v}^{(attribute)} + \sum_t \alpha^{v,t}\mathbf{v}^{(neighbor),t}, \tag{5}$$

where $\mathbf{v} \in \mathbb{R}^d$, $\alpha^{v,*}$ indicates the importance of different embeddings, $\mathbf{v}^{(attribute)}$ is the textual attribute embedding of v, and $\mathbf{v}^{(neighbor),t}$ is the type-based aggregated embedding obtained from Eq.(2). Here the importance weight can be written as:

$$\alpha^{v,t} = \frac{exp(LeakyReLU(\mathbf{a}^T)[\mathbf{v}^{(attribute)} \oplus \mathbf{v}^{(neighbor),t}])}{\sum_u exp(LeakyReLU(\mathbf{a}^T)[\mathbf{u}^{(attribute)} \oplus \mathbf{u}^{(neighbor),t}])}, \tag{6}$$

where u is a node in set $\mathcal{N}_t(v)$, $LeakyReLU$ denotes leaky version of a Rectified Linear Unit, $\mathbf{a} \in \mathbb{R}^{2d}$ is the attention parameter.

We use the embedding of the local content node (e.g., the author's paper/message) as the behavior's NetC embedding $\mathbf{h}_{NetC(b)}$.

3.3 The Interactive Hierarchical Attention Mechanism

In location attention, the weights were calculated separately on each type of context. However, it is important to model the interactions between the contexts (as described in the Introduction). We propose a new mechanism. First of all, it models pairing effects of contexts – they are embeddings generated by weighting and summing embedding of two types of contexts:

$$\mathbf{h}_{C(b)_i,C(b)_j} = \gamma_{i,j} \cdot \mathbf{h}_{C(b)_i} + (1 - \gamma_{i,j}) \cdot \mathbf{h}_{C(b)_j}, \tag{7}$$

where $C(b)_i$ is the i-th type of b's behavioral context, such as LC(b), NC-f(b), NC-l(b), RC(b), and NetC(b); $\gamma_{i,j}$ is the weight of $C(b)_i$ over $C(b)_j$. We learn the weight $\gamma_{i,j}$ through a multilayer perceptron with one hidden layer and the ReLU activation function:

$$\gamma_{i,j} = \sigma(\mathbf{x}^T ReLU(\mathbf{W}_{i,j}(\mathbf{h}_{C(b)_i} \oplus \mathbf{h}_{C(b)_j}) + c_{i,j}), \tag{8}$$

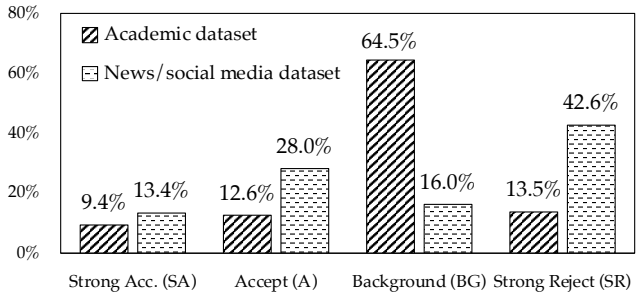


Figure 4: Distributions of intention types on two datasets.

where $\mathbf{x}$, $\mathbf{W}_{i,j}$, and $c_{i,j}$ are model parameters, σ is a sigmoid function. Then the concatenated embedding vector is

$$\mathbf{h}_b = \mathbf{h}_{LC(b),NCf(b)} \oplus \mathbf{h}_{LC(b),NCl(b)} \oplus \mathbf{h}_{LC(b),NetC(b)} \oplus \mathbf{h}_{LC(b),RC(b)} \oplus \mathbf{h}_{NCf(b),NCl(b)} \oplus \mathbf{h}_{RC(b),NetC(b)} \in \mathbb{R}^{6d}. \quad (9)$$

Then we apply a hierarchical attention on the behavior's concatenated embedding $\mathbf{h}_b$:

$$\hat{\mathbf{h}}_b = \sum_i \alpha_i \cdot (\mathbf{h}_b)_i, \quad (10)$$

$$\alpha_i = \frac{exp(\beta_i)}{\sum_i exp(\beta_i)}, \beta_i = f((\mathbf{h}_b)_i), \quad (11)$$

where α_i is the attention weight for the i-th value in $\mathbf{h}_b$, and $f(\cdot)$ is a multilayer perceptron function.

Then we can write the optimization function (i.e., training loss) as a negative likelihood of predicting correct intention labels of behaviors in the training data:

$$\mathcal{L} = -\sum_{b \in \mathcal{B}} \sigma(\mathbf{W} \cdot \hat{\mathbf{h}}_b + c), \quad (12)$$

where $\mathcal{B}$ is the data set of behaviors, $\mathbf{W}$ and c are parameters, $\hat{\mathbf{h}}_b$ can be obtained from Eq.(10), and $\sigma(\cdot)$ is the softmax function.

4 EXPERIMENTS

In this section, we first introduce our datasets and experimental settings. Then we present experimental results to demonstrate the effectiveness of the proposed interactive hierarchical attention mechanism for identifying referential intention.

4.1 Datasets in Two Different Domains

Academic dataset. We chose two popular research topics, *network embedding* and *language model*. We manually selected, downloaded, and annotated 20 papers for each topic. Most of the 40 papers were published in WWW, KDD, ICDM, AAAI, NeurIPS, and ACL. We collected paper information such as title, author list, venue, year, and full text. Here a referential behavior is a citation, e.g., "[22]" or "(Mikolov et al. 2013)". We labelled 1,565 citations in total.

News/social media dataset. We crawled 297 news articles and retweets about 20 political events in the United States in 2019. Each item has author, time, content, and referred contents (such as hyperlinks in news and original tweet of retweets). Retweets had their original tweets on the same page; hyperlinked texts in news were further crawled. We labelled 401 referential behaviors.

Data labelling. Three domain experts (in the two specific research topics) were recruited to label the data. Only when all labels were consistent, the data object (i.e., referential behavior and intention label) was included in the final datasets. So we have **1497** (95.7%) and **368** (91.8%) data objects in academic dataset and news/social media dataset, respectively. The consistency ratios are high (above 90%) showing the existence of behavioral intention. The intention can be identified after a long and careful investigation.

Statistics. Figure 4 presents the percentages of intention label types in the two datasets. In academic data, more than half of the citations were used for background description; around 1/5 (9.4%+12.6%) were used as the basis of the proposed method/idea; and around 1/8 (13.5%) had weak points that needed to be addressed. The numbers are consistent with commonsense. In news/social media data, 42.6% were holding a different opinion from the original news/tweet. Expressing a different (not the same) opinion is one of the common reasons that journalists write articles and people post messages.

4.2 Experimental Settings

Evaluation metrics. As it is a standard multi-class classification task, firstly, for each type of classes $s \in \mathcal{S}$, we calculate Precision and Recall, and report *Avg. Precision* and *Avg. Recall*. Secondly, we calculate the F1 score which is the harmonic average of the precision and recall. We use *Micro F1* which globally counts the TPs, FNs, FPs, and TNs. In our case, because all the concepts were assigned to exactly one class in the ground truth, the *Micro F1* is the same as *Accuracy*. We also use *Macro F1* which is the unweighted mean of the F1 scores per type of classes. Moreover, we plot the *Precision-recall curve* per type of classes. We calculate the *Area under the Receiver Operating Characteristic Curve (AUC)* for evaluation. For all the metrics above, higher score means better performance.

Competitive methods. We will compare our method with competitive baselines as follows:

- CiteFrame (2018) [14]: This system proposed pattern-based, topic-based, and prototypical argument features to model citations. We used the same set of features and classifiers for intention classification on the academic dataset.
- CiteFunc (2019) [19]: This was a multi-task learning method based on GloVe embeddings and CNN architecture. The two tasks were citation function and provenance classifications.
- SecClass (2019) [6]: This work was to classify citation intent into three categories: background, use of method, and result comparison. So the section, not semantics, would make the most significant impact.

Besides the three recent baselines, we implemented as many as 12 variants of our proposed method to analyze the following points:

- We will compare M1–M4 to analyze the importance of RC and NetC on both academia and social media datasets.
- We will compare M4–M8 to analyze the importance of NC using ablation studies on academia data.
- We will compare 3 different attention mechanisms (local attention only, hierarchical attention without interaction, and the proposed *interactive hierarchical attention*) on M8 that uses full contexts, with both datasets.

	Academic dataset (%)					News/social media dataset (%)				
	Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1	AUC
CiteFrame (2018) [14]	63.59	54.95	52.66	52.72	71.61	-	-	-	-	-
CiteFunc (2019) [19]	71.74	63.22	53.61	57.12	74.83	56.25	57.01	56.11	53.67	73.64
SecClass (2019) [6]	79.35	67.71	68.43	67.86	83.04	61.25	61.87	64.07	62.72	78.57
Ours (**M8-IHA**)	**82.21**	**77.01**	**70.14**	**73.14**	**89.66**	**68.75**	**70.36**	**72.10**	**70.72**	**82.31**

Table 3: Our proposed method significantly outperform existing methods on both datasets for identifying referential intention.

	Contexts				Academic dataset (%)					News/social media dataset (%)				
	LC	NC	RC	NetC	Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1	AUC
M1	✔				77.16	66.94	63.39	64.77	84.42	61.00	60.30	62.05	62.08	74.45
M2	✔		✔		79.35	71.35	62.10	65.59	84.60	63.75	64.51	65.24	64.46	81.40
M3	✔			✔	79.62	75.34	60.44	65.33	84.27	65.00	67.26	66.83	66.79	80.02
M4	✔		✔	✔	79.89	74.11	64.90	68.45	84.46	67.75	67.33	69.97	67.94	81.75
M8(LA)	✔	✔	✔	✔	81.01	74.26	68.35	70.84	89.08	67.75	67.33	69.97	67.94	81.75
M8-HA	✔	✔	✔	✔	81.79	75.06	69.08	71.57	89.35	67.50	69.91	68.83	68.73	82.25
M8-IHA	✔	✔	✔	✔	**82.21**	**77.01**	**70.14**	**73.14**	**89.66**	**68.75**	**70.36**	**72.10**	**70.72**	**82.31**

Table 4: We compared different variants and attention mechanisms, i.e., local attention only, hierarchical attention without interaction, and the proposed *interactive hierarchical attention* (IHA), on M8 both datasets. Because there was no neighboring context (NC) in news/social media dataset, our M8 is actually M4 (M8-LA equals to M4). M8-IHA performs the best.

	Contexts				Academic dataset (%)				
	LC	NC	RC	NetC	Acc.	Prec.	Rec.	F1.	AUC
M4	✔		✔	✔	79.89	74.11	64.90	68.45	84.46
M5	✔	✔			78.53	69.14	69.21	68.46	87.80
M6	✔	✔		✔	80.77	73.51	67.92	70.17	88.76
M7	✔	✔	✔		**81.52**	72.35	**69.94**	70.54	**89.50**
M8	✔	✔	✔	✔	81.01	**74.26**	68.35	**70.84**	89.08

Table 5: Ablation studies: Neighboring context (NC) and referred content (RC) provide semantic contexts of the citation. So they are important for identifying citation intents.

- We will compare M5, M9, and M10 to learn the impact of NC-former and NC-latter on academia data.

Note that except M8-HA and M8-IHA, all the variants use local intention only, for fair comparison with the three outside baselines.

4.3 Experimental Results

Overall performance. Experiments on the two real datasets demonstrate the effectiveness of our proposed method M8-IHA: F1 was improved relatively by **+12.92%** on the academic dataset and by **+13.92%** on the news/social media dataset over pure M1 (LC only). On academic dataset, integrating rich textual contexts and their interactions can achieve satisfactory performance. NetC shows a great contribution to intention classification on news/social media.

4.3.1 Results on academic dataset.

Comparison with baselines. Our purposed methods (from M4-, including the best M8-IHA) outperform all baseline methods, as shown in Table 3. Our method improves accuracy relatively by **+3.73%**, precision by **+13.73%**, recall by **+2.50%**, and F1 score by **+7.78%**, respectively comparing with SecClass(2019). Our method models heterogeneous contexts as well as their interactions of rich semantics and structures that are strongly related to referential intention. The baselines were not able to extract and learn the important information.

	Contexts			Academic dataset (%)				
	LC	NC-f	NC-l	Acc.	Pre.	Rec.	F1.	AUC
M5	✔	✔	✔	78.52	69.14	**69.21**	**68.46**	**87.80**
M9	✔	✔		78.26	**73.22**	65.14	68.19	87.49
M10	✔		✔	**78.53**	69.16	67.98	68.11	87.56

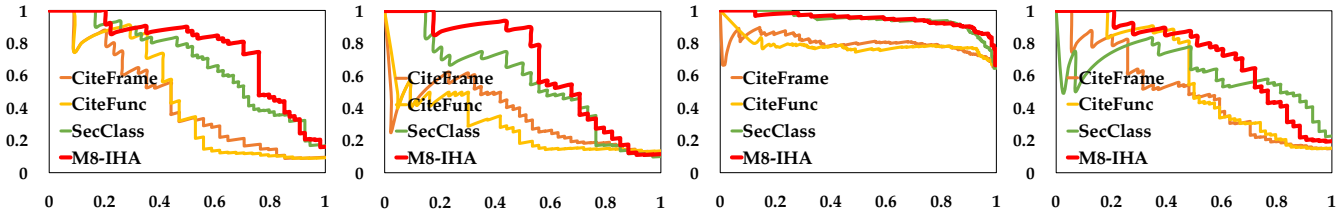
Table 6: Comparing the impact of *NC-f*ormer and *NC-l*atter. NC-l may hurt precision but significantly improve recall.

Analyzing impact of heterogeneous contexts. Table 4 compares M1–M4 for analyzing the impact of referred content (RC) and network context (NetC) on identifying referential intention in academic and social media datasets. Compared with M1, adding RC (M2) and NetC (M3) improves F1 score relatively by +1.22% and +0.86%, respectively. When having both contexts (M4), F1 can be further improved by **+4.36%** and **+4.78%** comparing with M2 and M3, respectively. On academic dataset, the improvements by RC (M2) and NetC (M3) are comparable; on social media, NetC (M3) makes greater improvement than RC (M2).

Table 4 compares M4–M8 that all adopt the local attention only but use different combinations of contexts. This is literally ablation study on academic dataset. By using all types of contexts, M8 delivers the best performance on F1 score. Compared with M1 (LC only) , M8 improves accuracy relatively by **+6.54%**, precision by **+15.04%**, recall by **+10.65%**, and F1 score by **+12.92%**.

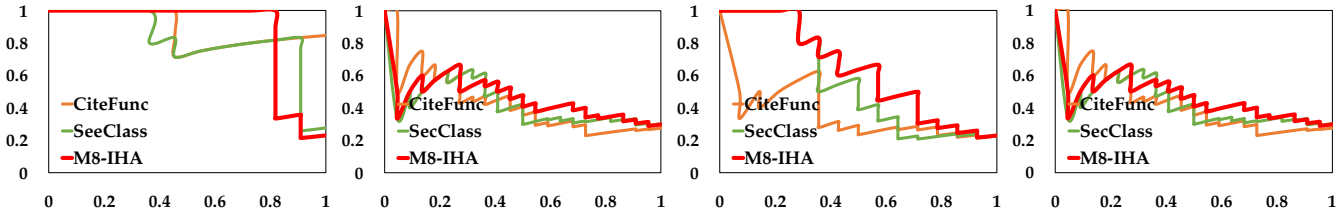
Table 6 compares M5, M9, and M10. Both neighboring contexts (NC-f & NC-l) can improve model effectiveness, if combined with RC and NetC (see M6–M8); however, it improves only +0.39% and +0.51% (no more than 1%) compared with M9 and M10. We observe that NC-l can significantly improve recall but may hurt precision.

Analyzing attention mechanisms. Table 4 also compares different attention mechanisms, i.e., local attention only (M8 (LA)),



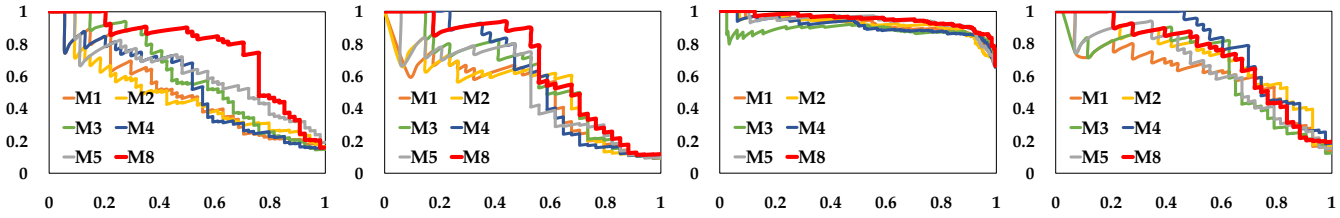
(a) Strong Accept (SA) (b) Accept (A) (c) Background (BG) (d) Strong Reject (SR)

Figure 5: Precision-recall curves of M8-IHA (the best of ours) with baseline methods on the *academic dataset.*



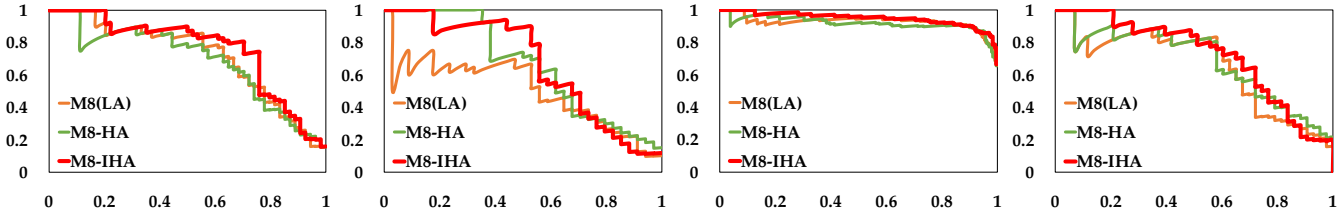
(a) Strong Accept (SA) (b) Accept (A) (c) Background (BG) (d) Strong Reject (SR)

Figure 6: Precision-recall curves of M8-IHA (the best of ours) with baseline methods on the *news/social media dataset.*



(a) Strong Accept (SA) (b) Accept (A) (c) Background (BG) (d) Strong Reject (SR)

Figure 7: All types of contexts matter: Precision-recall curves of M1–M4, M5, and M8 on the *academic dataset.*



(a) Strong Accept (SA) (b) Accept (A) (c) Background (BG) (d) Strong Reject (SR)

Figure 8: Interactive Hierarchical Attention wins: Precision-recall curves of M8 variants on the *academic dataset.*

hierarchical attention without interaction (M8-HA), and the proposed interactive hierarchical attention (M8-IHA) on M8 that uses all types of contexts with both datasets. M8-IHA achieves a better performance than the other two models. We observe that neural frameworks of *flat* local attention or *pure* hierarchical attention mechanisms cannot perform well. The interactive attention design can significantly improve the performance. It includes multiple weighted pooling modules. Each pooling module integrates outputs of two types of contexts with local attention. Clearly, multiple types of contexts have interaction with each other: (1) LC and NC came originally from the same document and form semantic ordering; (2) RC and LC were tightly related to the same topic; (3) RC and LC were produced by a pair of linked nodes in NetC, and so on. Our proposed IHA mechanism can better distill useful information in the multiple types of contexts. Compared with LA and HA, IHA improves F1 score relatively by **+3.25%** and **+2.19%**, respectively.

We investigated the interactions in depth: The interaction between generated content (GC = LC & NC) and referred context (RC) makes greater impact than the interaction between network context (NetC) and other type of contexts. Both GC and RC are textual contexts and certainly have semantic relationships between each other. NetC and GC have different modalities (network and text, respectively) and are embedded by different types of encoders.

Analyzing performances on intention types. Figure 5 shows precision-recall curves of the best of our proposed methods (M8-IHA) and three competitive baselines on the academic dataset. We observe that our proposed method consistently performs better than the baselines on polar intentions (i.e., Strong Accept, Accept, and Strong Reject). On the type of background (BG), which is the

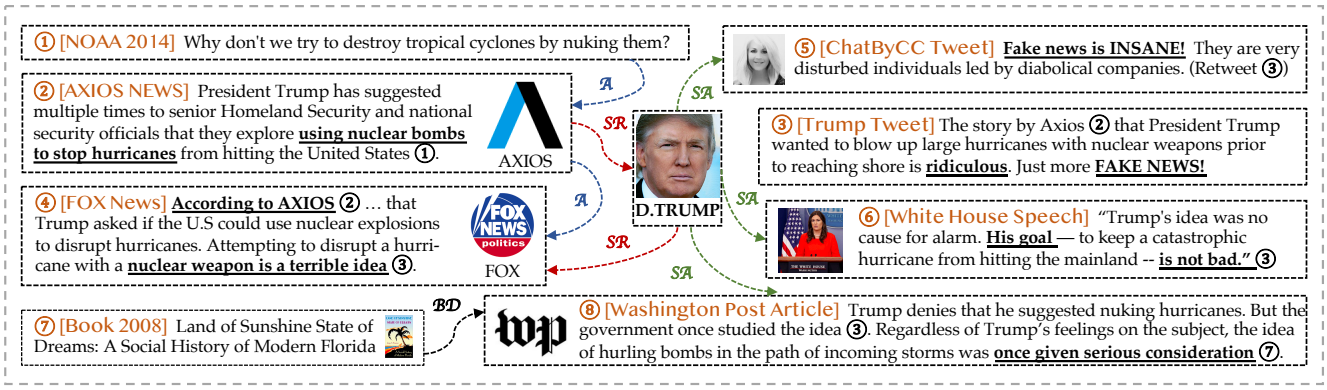


Figure 9: A case of referential behaviors on news and social media. The serial numbers ① – ⑧ follow the chronological order of the generated contents. (③ was posted by D. Trump. It rejected a new article; it was accepted by multiple messages; and it was also rejected by multiple articles.) The directed links represent the referential behaviors. The link attributes indicate the predicted referential intentions. Some contents were not fully quoted due to space limitation. Our proposed method provides useful information from the perspective of behavioral sciences for knowledge graph construction and behavior data mining.

most frequent intention type of the imbalanced academic dataset, our method and SecClass have comparable performances. They are both close to perfect. Compared to polar intentions, BG is relatively easy to be identified. The "background" referential behaviors are often in the section of "Related Work", so the section feature is significant on identifying the BG intention.

Figure 7 shows precision-recall curves of M1–M4, M5, and M8 on the academic dataset. On the referential intentions of Strong Accept and Accept, M8 performs the best; on Background, all methods are comparablely good; however, on the intention of Strong Reject, M4 performs the best – So, the neighboring contexts (NC-former & NC-latter) may have negative effect on identifying the negative intention. The reason is that if the intention is Strong Reject, the local context (LC) has pointed out the weak point of referred content (RC); LC and RC can fully reflect the comparative information. Neighboring sentences may have been talking about the author's work which are likely to include positive words of the work's contributions. So, NC may confuse the model.

Figure 8 shows precision-recall curves of M8's variants on different attention mechanisms, i.e., local attention only, hierarchical attention without interaction, and the proposed interactive hierarchical attention (IHA). The IHA mechanism is more effective on identifying Accept and Strong Accept.

4.3.2 *Results on news/social media dataset.*

Comparison with baselines. Because there was no neighboring context (NC) in this dataset, **our M8-IHA is actually M4-IHA** which uses LC, RC, and NetC as well as a partial interactive hierarchical attention mechanism. It significantly outperforms all baselines. Compared with the best baseline, our purposed method improves accuracy relatively by **+12.24%**, precision by **+13.72%**, recall by **+12.53%**, and F1 score by **+12.76%**, respectively.

Analyzing impact of heterogeneous contexts. As shown in Table 4, the network context (NetC) improves the F1 score relatively by **+7.59%**. With all types of contexts, M4 achieves the best performance: it improves accuracy relatively by **+12.70%**, precision by **+16.62%**, recall by **+16.24%**, and F1 score by **+13.92%**, respectively, compared to local context (LC) only. We have consistent observations with the ones on the academic dataset: by combining all referential contexts, we can have a more satisfactory performance. NetC makes a greater contribution than referred context (RC). User generated contents have multiple kinds of expressions reflecting complex sentiments (e.g., irony). Such expressions make it more difficult to identify referential intentions on the news/social media dataset than on the academic dataset.

Analyzing performances on intention types. Figure 6 shows precision-recall curves of the best of our proposed methods (M8-IHA) and three competitive baselines on the news/social media dataset. Our proposed method achieves better results on the prediction of each intention label. On the news/social media dataset, Strong Accept usually contains literal praise, so predicting SA achieves the highest accuracy.

4.4 Case Study on News/Social Media Data

Figure 9 presents a news event including multiple referential behaviors from different sources. The event originated from a news report from AXIOS that "*President Donald Trump suggested to explore using nuclear bombs to stop hurricanes from hitting the United States*". We use this example to analyze results of M1–M4 methods.

②→①: **Accept (A).** AXIOS news said that using nuclear weapons to destroy the storms has been so persistent by the National Oceanic and Atmospheric Administration (NOAA) to support its claim. M1 and M3 predicted *BackGround* while M2 and M4 predicted *Accept*. The referred content (RC) could distinguish ②'s referential intention. On one hand, NOAA gave detailed explanation to not trying to destroy tropical cyclones by nuking them so the author's stance in ② is positive. On the other hand, the official interpretation of NOAA has certain credibility, so reader's stance of ① is assumed to be positive. Therefore, *Accept* is more reasonable than *BackGround*.

⑤→③: **Strong Accept (SA).** CC's tweet directly supported Donald Trump's claim, pointing out that AXIOS reported fake news. M1 and M3 predicted *Strong Reject* while M2 and M4 predicted *Strong*

Accept. The referred content (Trump's tweet) helped distinguish ⑤'s referential intention. CC's stance in ⑤ is clearly positive to ③ (actually negative to AXIOS news ②). Therefore, *Strong Accept* is more reasonable than *Strong Reject*.

⑧→③: Strong Accept (SA). The Washington Post supported Donald Trump's idea through a historical event to indicate nuking hurricanes is actually worth considering. M1 and M2 predicted *Accept*, while M3 and M4 predicted *Strong Accept*. The network context (NetC) helps the method better identify ⑧'s referential intention. On one hand, the Washington post's stance in ⑧ is positive because it found clues in history to persuade people Trump's advice is feasible. On the other hand, Trump often reported fake news on his Twitter so reader's stance is neutral and even less convinced. Therefore, *Strong Accept* is more reasonable than other intention labels such as *Accept*, *Background*, and *Strong Reject*.

⑧→⑦: Background (BG). The Washington Post in ⑧ introduced a historical event from a book ⑦ that after the United States dropped the first atomic bombs on Hiroshima and Nagasaki in 1945, Floridians started wondering if the same destructive force that had ushered in the end of World War II could also be leveraged to protect beachfront property. M2 and M3 predicted *Accept*, while M4 predicted *BackGround*. On one hand, the story in the book ⑦ is not directly related to Trump's idea. This Washington Post's article does not have a positive or negative attitude to the story in the book. Therefore, *BackGround* is more reasonable than other intention according to our proposed schema.

④→③: Strong Reject (SR). This doesn't need more explanation. All methods M1–M4 predicted the correct intention label.

5 RELATED WORK

In this section, we review existing work related to our study including stance detection and intent classification schema and methods.

5.1 Stance Detection

Stance detection is to detect the stance ("Favor" or "Against") expressed in text towards a specific target. It is an emerging problem in sentiment analysis. Previous work in stance detection mostly focused on debates [12, 24] or student essays [10]. There is a growing interest in performing stance classification on microblogs such as Twitter and Weibo [9, 17, 28]. Du *et al.* incorporated target-specific information into stance classification by following an attention mechanism [9]. Zhou *et al.* proposed an attention mechanism in the bidirectional GRU-CNN structure to perform target-specific stance detection on tweets [30]. Dey *et al.* proposed a two-phase LSTM based model with attention [7] and Wei *et al.* proposed an end-to-end neural memory model via target and tweet interactions [25]. By considering the dependency of related targets, Sobhani *et al.* introduced a multitarget stance detection task and proposed an attentive encoder-decoder network to capture the dependencies among stance labels regarding multiple targets [18]. Wei *et al.* proposed a dynamic memory-augmented network that utilized a shared external memory to capture and store multi-targets stance indicative clues [26]. Siddiqua *et al.* proposed a neural ensemble model that adopted the strengths of two LSTMs to learn long-term dependencies, where each module coupled with an attention mechanism that amplifies the contribution of important elements in the final representation [17]. However, their goal was to analyze the stance towards a specific target mentioned in text. Our work is different – it is to identify referential intention when an object refers to a specific claim from the other object.

5.2 Intent Classification Schema

There has been a wide line of research on classifying citation intents [2, 3, 8, 11, 20, 27]. Volenzuela *et al.* introduced a novel task of identifying important citations in scholarly literature, defined with two categories: *important* vs. *non-important* [21]. Jurgens *et al.* captured broad thematic functions a citation can serve in the discourse [13, 14]. They defined six categories based on functions of citation content on the topic, such as providing background knowledge and describing motivation. Later, Cohan *et al.* proposed a concise annotation schema including three categories that is useful for navigating research topics and machine reading of scientific papers [6]. However, these schema were based on natural language understanding to classify sentences, which ignores characteristics of author, generated content, referred content, and network of contents.

5.3 Intent Classification Methods

Early work in citation intent classification was mainly based on manual analysis [11, 15] or rule-based systems [11, 16]. They lacked generalizability and scalability. To address the issues, Abu-Jbara *et al.* proposed a supervised method for identifying citation text and analyzing it to determine the purpose and the polarity of citation [1]. Volenzuela *et al.* extracted 12 predefined features from citation and fed them into a random forest classifier [21]. Jurgens *et al.* expanded all pre-existing feature-based efforts on citation intent classification by proposing a comprehensive set of engineered features, including bootstrapped patterns, topic modeling, dependency-based, and metadata features for the task [13, 14]. However, their methods are based primarily on extracting predefined pattern-based, topic-based, or prototypical argument features. With widespread use of neural models, researchers have investigated many neural methods for this task. Su *et al.* applied a convolutional neural network (CNN) to classify both citation function and provenance, which surpassed the performance of rich-feature based baselines [19]. Cohan *et al.* adopted an attention-based BiLSTM model and proposed structural scaffolds to incorporate knowledge into citations from scientific papers structures for effective classification of citation intents [6]. Nevertheless, these methods only learned sentence representation, which depends solely on syntax and semantics of sentences.

6 CONCLUSIONS

In this work, we proposed to identify the intention of referential behaviors. We adopted a theory in sociology to develop a schema of four types of intentions. We propose a new neural framework with Interactive Hierarchical Attention (IHA) to identify the intention of referential behavior by properly aggregating the heterogeneous contexts. Experiments demonstrate that the proposed method can effectively identify the type of intention of citing behaviors (on academic data) and retweeting behaviors (on Twitter).

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant IIS-1849816.

REFERENCES

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL)*. 596–606.

[2] Shashank Agarwal, Lisha Choubey, and Hong Yu. 2010. Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium Proceedings*, Vol. 2010. American Medical Informatics Association, 11.

[3] Tanzila Ahmed, Ben Johnson, Charles Oppenheim, and Catherine Peck. 2004. Highly cited old papers and the reasons why they continue to be cited. Part II., The 1953 Watson and Crick article on the structure of DNA. *Scientometrics* 61, 2 (2004), 147–156.

[4] Olga Amsterdamska and Loet Leydesdorff. 1989. Citations: Indicators of significance? *Scientometrics* 15, 5-6 (1989), 449–471.

[5] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2016. Deep neural networks for learning graph representations. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*.

[6] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL)* (2019), 3586–3596.

[7] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention. In *European Conference on Information Retrieval (ECIR)*. Springer, 529–536.

[8] Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th international joint conference on natural language processing (IJCAI)*. 623–631.

[9] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence (IJCAI).

[10] Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In *The Twenty-Seventh International Flairs Conference*.

[11] Mark Garzone and Robert E Mercer. 2000. Towards an automated citation classifier. In *Conference of the canadian society for computational studies of intelligence*. Springer, 337–346.

[12] Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*. 1348–1356.

[13] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2016. Citation classification for behavioral analysis of a scientific field. *arXiv preprint arXiv:1609.00435* (2016).

[14] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics (TACL)* 6 (2018), 391–406.

[15] Michael J Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social studies of science* 5, 1 (1975), 86–92.

[16] Son Bao Pham and Achim Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 759–771.

[17] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet Stance Detection Using an Attention based Neural Ensemble Model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL)*. 1868–1873.

[18] Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 551–557.

[19] Xuan Su, Animesh Prasad, Min-Yen Kan, and Kazunari Sugiyama. 2019. Neural multi-task learning for citation function and provenance. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 394–395.

[20] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, 103–110.

[21] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems (NIPS)*. 5998–6008.

[23] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations (ICLR)*.

[24] Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL)*. Association for Computational Linguistics, 592–596.

[25] Penghui Wei, Junjie Lin, and Wenji Mao. 2018. Multi-target stance detection via a dynamic memory-augmented network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*. ACM, 1229–1232.

[26] Penghui Wei, Wenji Mao, and Daniel Zeng. 2018. A target-guided neural memory model for stance detection in Twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[27] Howard D White. 2004. Citation analysis and discourse analysis revisited. *Applied linguistics* 25, 1 (2004), 89–116.

[28] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications*. Springer, 907–916.

[29] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 793–803.

[30] Yiwei Zhou, Alexandra I Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*. Springer, 18–32.