

# Deep Learning of Knowledge Graph Embeddings for Semantic Parsing of Twitter Dialogs

Larry Heck  
Microsoft Research  
larry.heck@ieee.org

Hongzhao Huang  
Rensselaer Polytechnic Institute  
huangh9@rpi.edu

**Abstract**—This paper presents a novel method to learn neural knowledge graph embeddings. The embeddings are used to compute semantic relatedness in a coherence-based semantic parser. The approach learns embeddings directly from structured knowledge representations. A deep neural network approach known as Deep Structured Semantic Modeling (DSSM) is used to scale the approach to learn neural embeddings for all of the concepts (pages) of Wikipedia. Experiments on Twitter dialogs show a 23.6% reduction in semantic parsing errors compared to the state-of-the-art unsupervised approach.

**Index Terms**—deep learning, semantic parsing, Twitter, dialog

## I. INTRODUCTION

With the goal of teaching machines to understand human conversations, one of the most fundamental components of a conversational understanding system is the semantic parser. Conversational semantic parsers map natural language (NL) to a formal representation of meaning, typically defined by the intent of the user and the associated arguments of the intent (slots or concepts) [1].

Considerable advancements in semantic parsing have been made possible by the availability of massive volumes of data from web search engines. The data is composed of unstructured and semi-structured queries and documents from the web. User clicks on documents provide a (weak) semantic annotation of the query. Our recent work has shown significant progress in leveraging these weak click labels to semantically parse conversational queries [2], [3], [4], [5], [6].

With the recent emergence of very large-scale semantic knowledge graphs (KGs) [7], it is now possible to add structure to the machine learning procedures developed above. Specifically, we have developed methods to enrich KGs with automatically annotated training data through unsupervised data mining methods. The methods have been applied to queries (single turns in a conversation) [8], [9], [10], [11], [12], [13], [14], [15] as well as over multiple queries in a conversation [16].

Recent work in [17], [18] showed that a deep neural network (DNN) can be used to directly encode KGs for semantic parsing. By leveraging our prior work, this paper extends these KG methods in several ways. While the approach in [17] is single-relation Question Answering, our approach is large-scale multi-concept (entity, relation, fact) open domain semantic parsing. Our approach is web-scale, learning neural embeddings for all the concepts of Wikipedia in the open

source Freebase KG [7]. Also, while the other approaches rely on supervised training, our approach is unsupervised. Also, our approach is applied to the Wikification (concept linking) of Twitter Tweets as opposed to the Question Answering problem. And finally, we have extended the KG-based semantic parsing methods to multi-turn (Twitter) dialogs.

## II. TASK DEFINITION

This paper focuses on the task of semantically parsing human dialogs (spoken or text). With the proliferation of mobile devices and services/Apps such as Twitter, Skype, and Facebook, the dominant mode of human dialog is being redefined. We have entered a new era where an ever increasing number of people are continuously connected through digital conversations composed of a sequence of short messages. The availability of large volumes of data for scientific study from these conversations (e.g., Twitter) presents a new opportunity for the semantic parsing community to explore the potential of automated understanding of human dialog.

Figure 1 shows a series of tweets separated by punctuation (or a multi-sentence tweet). We seek to semantically parse the tweets; specifically, identify the unambiguous Wikipedia concepts (entities, relations, and facts) present in the utterance. The figure shows the concept mentions in bold, including “Hawks”, “Fans”, and “slump”. Concept mentions are defined as a natural language surface form phrase referring to a concept. The unambiguous concept in this task is defined as a Wikipedia page or URL (e.g., [http://en.wikipedia.org/wiki/slump\(sports\)](http://en.wikipedia.org/wiki/slump(sports))). Concept linking is completed with no constraints on the domain or topic of conversation. This task is often referred to as *large-scale open domain entity linking* [19].

## III. COHERENCE AND SEMANTIC RELATEDNESS

With multiple dialog turns in a Twitter conversation, one can exploit the topical coherence and semantic relatedness between concept mentions. While a single mention may be ambiguous, multiple mentions of the same concept in different grammatical context over a dialog often provide additional evidence of the unambiguous (Wikipedia) concept. In addition, mentions of other semantically related concepts provide additional disambiguation evidence.

Referring again to Figure 1, topical coherence is illustrated with the mention “Hawks”. This mention is repeated in the first



Fig. 1: Example of Twitter tweets and the concept linking task.

and third tweet/sentence. The combination of the mentions “Hawks”, “Fans”, and “slump” illustrate the importance of semantic relatedness. The mention “Hawks” may refer to the concept of an animal or the NBA team (Atlanta Hawks). The mention “Fans” may refer to the concept of people (sports fans) or a device for creating a current of air or breeze. The mention “slump” may refer to the concept in sports (a period when a player or team is not performing well) or a geological mass movement process of hill slope failure.

Exploiting topical coherence and semantic relatedness has resulted in improvements for Wikification on formal texts (e.g., News) [20], [21], [22], [23], [24]. For Twitter, however, the brevity of a single tweet typically does not provide enough topical context. Therefore, a method is required that captures semantic relatedness from multiple tweets.

The unsupervised method in this paper achieves this goal. It extends previous methods by (1) using a graph regularization method to combine the prior popularity of the concept with multi-turn (dialog) coherence, (2) leveraging large-scale knowledge graphs to learn neural KG embeddings, using these to compute semantic relatedness between mention-concept pairs. The graph regularization method uses a graph-based semi-supervised learning algorithm. Given a relational graph, the method regularizes a loss function of predicted mention-concept labels with a measure of label consistency over the graph. The graph regularization method jointly performs mention detection and disambiguation, and incorporates both local and global evidence from multiple tweets by detecting meta path-based semantic relations from social networks. A detailed description of the method is presented in [25].

In the next section, we detail our new approach to neural KG embeddings and describe how the embeddings can be combined with the graph regularization approach for semantic parsing of twitter dialogs.

#### IV. KNOWLEDGE GRAPH EMBEDDINGS

Most rule-based and statistical natural language processing (NLP) methods consider words as the lowest level atomic unit. Each word is represented as a binary “one-hot” vector, where the dimension of the vector is the size of the vocabulary. This representation of words is sparse and poorly captures the semantic relatedness between words, e.g., the AND operation

of “city” and “town” is zero (False). As a result, alternative feature representations of words are needed that are more compact and better capture semantic relatedness.

Deep neural networks have shown significant potential in creating compact feature representations for many applications - from image, speech, signal, and natural language processing. In the 1990s, deep neural network methods were developed to automatically discover new feature representations for robust speaker recognition [26], [27]. In this work, neural embeddings were discriminatively learned without supervision from raw acoustic features (large windows of filterbank spectral energies). The neural embeddings yielded significant improvements over previous state-of-the-art approaches. Compared to a top performing system on the 1998 NIST Speaker Recognition evaluations, the neural embeddings produced gains of greater than 28% in error rate reductions.

Recently, neural embeddings have been applied to text processing and used successfully to learn more effective representations of words. These representations, called *neural word embeddings*, combine vector space semantics with the prediction power of probabilistic models and yield dense vector representations. Neural embeddings for text-based language modeling and NLP applications were developed in [28], [29], [30]. These methods learn neural embeddings for a word from the other words that often occur in close proximity in documents (e.g., Wikipedia articles).

We seek to extend the word embedding methods to create *neural knowledge graph (KG) embeddings*: a dense, continuous-valued semantic vector representation of KG concepts. KG concepts often consist of multiple words (e.g., “Microsoft Research”, “James Cameron”, “Atlanta Hawks”). Previous extensions to word-based embeddings have typically represented multi-word concepts as algebraic combinations (addition) of word-level embeddings. While this approach can work for some concepts, it often introduces noise into the representation. For example adding the vector “Atlanta” (the city) to the vector “Hawks” (the bird) does not result in the vector for “Atlanta Hawks” (the basketball team).

Our approach learns neural embeddings of semantic concepts directly from the KG. For each concept, we identify the associated sub-graph of the KG and encode the knowledge as feature vectors. These features are used as input to a DNN that is trained to learn neural KG embeddings that represent the semantic relatedness between KG concepts.

For this work, we use the portion of the Freebase [7] that covers Wikipedia concepts (entities, relations, and facts) from a Wikipedia dump on May 3, 2013. The number of entities, entity types (e.g., person), and relations (e.g., directed by) used in this study are shown Wikipedia Table I.

TABLE I: Statistics of Freebase-Wikipedia Concepts

Knowledge Graph Element	Size
# Entities/Facts	4.12M
# Entity Types	1.57K
# Relations	3.17K

From this portion of Freebase, we generated several types of features to be used as input to the DNN. These features are shown in Figure 2. Given the small number of entity types and relations, we represent these features as a 1-of-V binary vector. With the relatively large number of entities, we efficiently represent the entity-based features by leveraging a method we developed in [31] called word hashing.

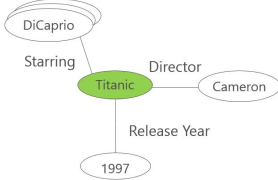
Knowledge	Representation	Example
Sub-Graph	Word hashing via letter tri-grams for ENTITIES  1-of-V vector for RELATIONS	
Entity Type	1-of-V vector	< 0...0...1...0 >

Fig. 2: Encoding of knowledge graph features

### A. Word Hashing

Word hashing aims to reduce the dimensionality of the bag-of-words term vectors. The specific approach we use is based on letter n-grams. As shown in Figure 3, given a word (cat), we first add start- and end-marks to the word (e.g., #cat#). Then, we break the word into letter n-grams (e.g., letter trigrams: #ca, cat, at#). Finally, the word is represented using a vector of letter n-grams.

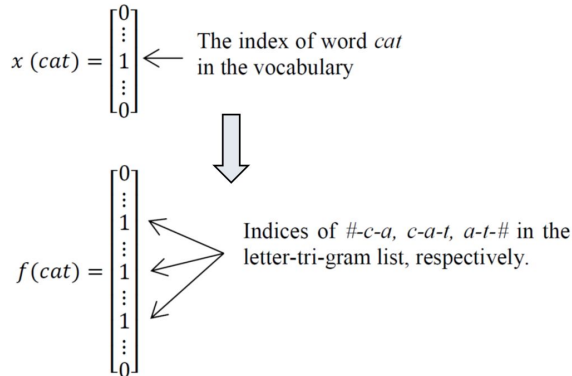


Fig. 3: Word hashing with letter tri-grams.

One potential problem of this method is collision, i.e., two different words could have the same letter n-gram vector representation. Table II shows some statistics of word hashing on two vocabularies. Compared with the original size of the one-hot vector, word hashing allows us to represent a sequence of words using a vector with much lower dimensionality. For example, each word of a 40K-word vocabulary can be represented by a 10,306-dimensional vector using letter trigrams,

giving a four-fold dimensionality reduction with few collisions. The reduction of dimensionality is even more significant when the technique is applied to a larger vocabulary. Each word in the 500K-word vocabulary can be represented by a 30,621 dimensional vector using letter trigrams, a reduction of 16-fold in dimensionality with a negligible collision rate of 0.0044

TABLE II: Word Hashing Collision Rate

Vocabulary Size	Observed Tri-letters in Vocabulary (unique)	Number of Collisions
40K	10306	2
500K	30621	22

### B. Deep Structured Semantic Models

To create the neural KG embeddings, the next step after generating the features from the KG is to input these features into a DNN and train the network to learn semantic relatedness between concepts. For this work, we employ the DNN method developed in [31] called ‘‘Deep Structured Semantic Models’’ (DSSM).

The architecture for the DSSM is shown in Figure 4. The feature generation is shown in the two bottom layers (Feature Vector and Word Hashing).  $E_i$  and  $E_j$  represent semantically related concepts (entities, relations, facts) and  $E_1, \dots, E_n$  represent non-related concepts (negative examples). Denoting  $x$  as the input term vector,  $y$  as the output vector,  $l_i, i = 1, \dots, N - 1$  as the intermediate hidden layers,  $W_i$  as the  $i$ -th weight matrix, and  $b_i$  as the  $i$ -th bias term, we have

$$\begin{aligned}
 l_1 &= W_1 x \\
 l_i &= f(W_i l_{i-1} + b_i), \quad i = 2, \dots, N - 1 \\
 y &= f(W_N l_{N-1} + b_N)
 \end{aligned} \tag{1}$$

where we use the  $\tanh$  as the activation function at the output layer and the hidden layers  $l_i, i = 2, \dots, N - 1$ :

$$f(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}. \tag{2}$$

The semantic relevance score between concepts  $E_i$  and  $E_j$  is then measured as:

$$R(E_i, E_j) = \cos(y_{E_i}, y_{E_j}) = \frac{y_{E_j}^T y_{E_i}}{\|y_{E_i}\| \|y_{E_j}\|} \tag{3}$$

where  $y_{E_i}$  and  $y_{E_j}$  are the neural embeddings of the KG concepts  $E_i$  and  $E_j$ , respectively. The semantic relatedness of two concepts is given by the KG as first-order related nodes (it can also be inferred from the co-occurrence of concepts on a given Wikipedia page). Given two semantically related concepts, the training procedure computes the posterior probability of concept  $E_j$  given  $E_i$  using a softmax function, as well as the probabilities for the unrelated concepts  $E_1, \dots, E_n$ . The DSSM is trained to maximize the likelihood of the related concepts given the features created across the KG. For a detailed description of the DSSM training procedure, we refer the reader to [31].

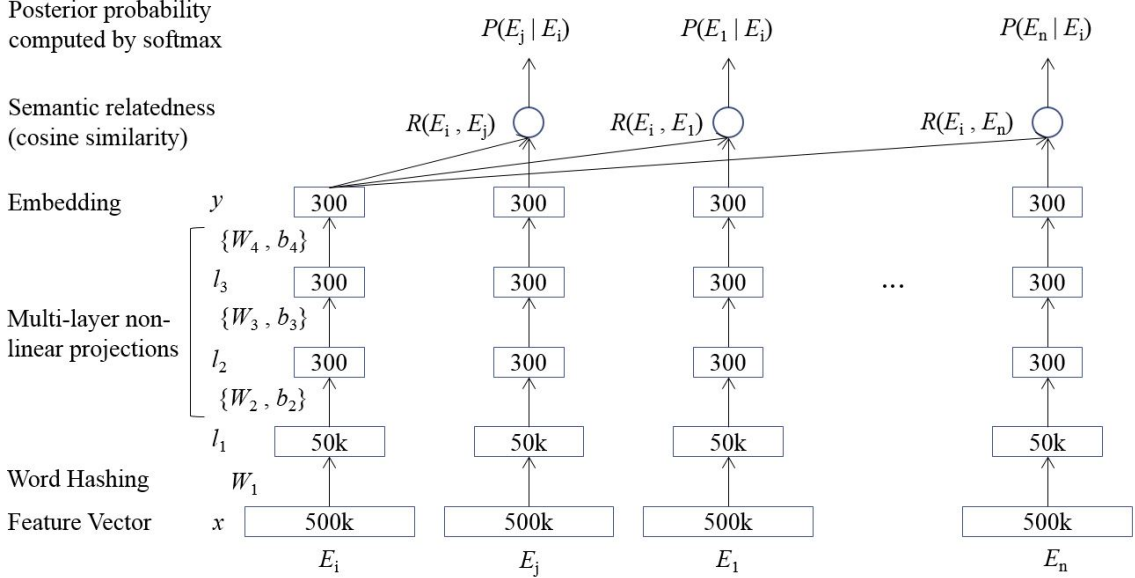


Fig. 4: DSSM architecture for learning neural knowledge graph embeddings

### C. Leveraging KG Embeddings with Graph Regularization

Given a KG embedding for each concept (Wikipedia page), we leverage the embeddings to improve the graph regularization approach described in Section III. The graph regularization approach generates candidate lists of coherent and semantically related concepts. In the original algorithm, the lists are ranked according to a semantic relatedness measure as proposed in [20], where semantic relatedness of two concepts  $c_i$  and  $c_j$  is computed as

$$SR(c_i, c_j) = 1 - \frac{\log \max(|C_i|, |C_j|) - \log |C_i \cap C_j|}{\log(|C|) - \log \min(|C_i|, |C_j|)} \quad (4)$$

where  $C_i$  and  $C_j$  are the set of incoming links to  $c_i$  and  $c_j$ , respectively. With the KG embeddings, we re-rank the candidate concept lists by replacing the measure of semantic relatedness of concepts with the DSSM posterior probability estimates, shown as  $P(E_j|E_i)$  in Figure 4.

## V. EXPERIMENTS

For our experiments we used a public data set (Meij et al., 2012) including 502 tweets posted by 28 verified users. The data set was annotated by two annotators. We used a Wikipedia dump on May 3, 2013, which included 30 million pages. A mention and concept pair  $\langle m, c \rangle$  was judged as correct if and only if  $m$  was linkable and  $c$  is the correct referent concept for  $m$ . To train the DSSM, we mined 20 million positive concept pairs. The negative training pairs were randomly sampled from across Wikipedia. Table III compares our new semantic parsing methods with the current state-of-the-art on this task called TagMe method by Ferragine et al. [32]. GraphRegu is our graph regularization approach described in Section III, and KG Embeddings is our new neural KG embedding approach. All results are error rates

of a hard decision using the top ranked concept candidate. For comparison, the state-of-the-art *supervised* method by Meij [33] has an error rate of 31.6%.

TABLE III: Deep conversational knowledge graph (Deep cKG) vs state-of-the-art TagMe.

Unsupervised Method	Error Rate	Reduction (rel.)
<i>Baseline (TagMe)</i>	38.1%	-
GraphRegu	35.7%	6.3%
+ KG Embeddings (Entities)	31.8%	16.5%
+ KG Embeddings (Relations)	30.0%	21.3%
+ KG Embeddings (Entity Types)	<b>29.1%</b>	<b>23.6%</b>

Using only our GraphRegu method reduces (improves) the error rate by 6.3% (relative) compared to the state-of-the-art TagMe system. Adding our neural KG embedding approach with first-order entities and facts in the sub-graph reduces the error rate by 16.5%. Including the KG relations and entity types yields a 21.3% and 23.6% error rate reduction. Our final system (unsupervised) has an error rate that is even lower than the state-of-the-art *supervised* method by Meij by 7.9% .

## VI. CONCLUSION

This paper presented a new unsupervised neural knowledge graph embedding model. The new model uses Deep Structured Semantic Modeling (DSSM) to learn the embeddings directly from large-scale knowledge graphs that cover all of Wikipedia. This paper also presented a semantic coherence-based approach for concept disambiguation across multiple dialog turns. When combined with the neural knowledge graph embeddings, the new approach yielded a 23.6% error reduction in the semantic parsing of Twitter dialogs.

## REFERENCES

- [1] G. Tur and R. D. Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. New York, NY: John Wiley and Sons, 2011.
- [2] D. Hakkani-Tür, L. Heck, and G. Tur, "Exploiting web search query click logs for utterance domain detection in spoken language understanding," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.
- [3] G. Tur, D. Hakkani-Tür, D. Hillard, and A. Celikyilmaz, "Towards unsupervised spoken language understanding: Exploiting query click logs for slot filling," in *Proceedings of Interspeech*, 2011.
- [4] D. Hakkani-Tür, G. Tur, L. Heck, A. Celikyilmaz, A. Fidler, D. Hillard, R. Iyer, and S. Parthasarathy, "Employing web search query click logs for multi-domain spoken language understanding," in *Proceedings of the IEEE ASRU Workshop*, 2011, pp. 419–424.
- [5] D. Hakkani-Tür, A. Celikyilmaz, L. Heck, and G. Tur, "A weakly-supervised approach for discovering new user intents from search query logs," in *Proceedings of Interspeech*, 2013, pp. 3780–3784.
- [6] L. Heck, D. Hakkani-Tür, M. Chinthakunta, G. Tur, R. Iyer, P. Parthasarathy, L. Stifelman, E. Shriberg, and A. Fidler, "Multi-modal conversational search and browse," in *Proceedings of the Workshop on Speech Language and Audio in Multimedia (SLAM)*, 2013, pp. 96–101.
- [7] "Freebase," <http://www.freebase.com>, 2014.
- [8] L. Heck, "The conversational web," in *Proceedings of the IEEE SLT Workshop*, Miami, FL, 2012.
- [9] L. Heck and D. Hakkani-Tür, "Exploiting the semantic web for unsupervised spoken language understanding," in *Proceedings of the IEEE SLT Workshop*, Miami, FL, 2012.
- [10] G. Tur, M. Jeong, Y.-Y. Wang, D. Hakkani-Tür, and L. Heck, "Exploiting semantic web for unsupervised statistical natural language semantic parsing," in *Proceedings of Interspeech*, 2012.
- [11] D. Hakkani-Tür, L. Heck, and G. Tur, "Using a knowledge graph and query click logs for unsupervised learning of relation detection," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.
- [12] L. Heck, D. Hakkani-Tür, and G. Tur, "Leveraging knowledge graphs for web-scale unsupervised semantic parsing," in *Proceedings of Interspeech*, 2013, pp. 1594–1598.
- [13] D. Hakkani-Tür, A. Celikyilmaz, L. Heck, G. Tur, and G. Zweig, "Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding," in *Proceedings of Interspeech*, 2014.
- [14] A. El-Kahky, D. Liu, R. Sarikaya, G. Tur, D. Hakkani-Tür, and L. Heck, "Extending domain coverage of language understanding systems via intent transfer between domains using knowledge graphs and search query click logs," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.
- [15] L. Heck, "Conversational knowledge graphs," Microsoft Research, Tech. Rep. MSR-TR-2014-70, 2014.
- [16] L. Wang, L. Heck, and D. Hakkani-Tür, "Leveraging semantic web search and browse sessions for multi-turn spoken dialog systems," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.
- [17] W. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proceedings of the ACL*, Baltimore, Maryland, June 2014.
- [18] A. Bordes, S. Chopra, and J. Weston, "Question answering with sub-graph embeddings," in *Proceedings of the EMNLP*, 2014.
- [19] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in *Proceedings of the EMNLP-CoNLL*, 2007.
- [20] D. Milne and I. Witten, "Learning to link with wikipedia," in *Proceedings of the 17th ACM CIKM*, October 2008, pp. 509–518.
- [21] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph-based method," in *Proceedings of SIGIR*, 2011.
- [22] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to wikipedia," in *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, 2011.
- [23] T. Cassidy, H. Ji, L. Ratinov, A. Zubiaga, and H. Huang, "Analysis and enhancement of wikification for microblogs with context expansion," in *Proceedings of COLING*, 2012, pp. 441–456.
- [24] X. Cheng and D. Roth, "Relational inference for wikification," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.
- [25] H. Huang, Y. Cao, X. Huang, H. Ji, and C. Lin, "Collective tweet wikification based on semi-supervised graph regularization," in *Proceedings of the ACL*, Baltimore, Maryland, June 2014.
- [26] Y. Konig, L. Heck, M. Weintraub, and K. Sönmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," in *Proceedings of RLA2C, ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications*, 1998, pp. 72–75.
- [27] L. Heck, Y. Konig, K. Sönmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, vol. 31, no. 2, pp. 181–192, 2000.
- [28] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proceedings of NIPS*, 2001.
- [29] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.
- [30] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.
- [31] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2013, pp. 2333–2338.
- [32] P. Ferragina and U. Scialla, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM CIKM*, 2010.
- [33] E. Meij, W. Weerkamp, and M. de Rijke, "Adding semantics to microblog posts," in *Proceedings of the 5th ACM International Conference on WSDM*, 2012.