7-2020

Interactive entity linking using entity-word representations

Pei Chi LO
Singapore Management University, pclo.2017@phdcs.smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

Citation

LO, Pei Chi and LIM, Ee-Peng. Interactive entity linking using entity-word representations. (2020). *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1801-1804.
Available at: https://ink.library.smu.edu.sg/sis_research/5269

Interactive Entity Linking Using Entity-Word Representations

Pei-Chi Lo
Singapore Management University
pclo.2017@phdcs.smu.edu.sg

Ee-Peng Lim
Singapore Management University
eplim@smu.edu.sg

Abstract

To leverage on entity and word semantics in entity linking, embedding models have been developed to represent entities, words and their context such that candidate entities for each mention can be determined and ranked accurately using their embeddings. In this paper, we leverage on human intelligence for embedding-based interactive entity linking. We adopt an active learning approach to select mentions for human annotation that can best improve entity linking accuracy at the same time updating the embedding model. We propose two mention selection strategies based on: (1) coherence of entities linked, and (2) contextual closeness of candidate entities with respect to mention. Our experiments show that our proposed interactive entity linking methods outperform their batch counterpart in all our experimented datasets with relatively small amount of human annotations.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

Interactive Entity Linking, Entity Representation, Knowledge Base

ACM Reference Format:
Pei-Chi Lo and Ee-Peng Lim. 2020. Interactive Entity Linking Using Entity-Word Representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401254

1 Introduction

Entity Linking (EL) is a task that links mentions of named entities from in an input text to entities in a knowledge base. As a core NLP task, EL has been included as an important component in many applications including question-answering and web search [2, 14]. To leverage on semantics derived from large data beyond the document containing the entity mention as well as the document covering the to-be-linked entity, researchers have recently proposed entity-word embedding models (e.g., Wikipedia2Vec[12]) to represent all entities and words in a common embedding space. In this way, each entity and word has its embedding enriched with additional semantics for EL. The accuracy of embedding-based EL however may suffer when the learned embeddings of entities do not match how their mentions are related to one another in the target document.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401254

EL can be performed in both batch and interactive modes. Batch EL links entity mentions in one pass. For difficult documents (e.g., a document containing multiple mentions which may be linked incorrectly), interactive EL allows human-in-the-loop to guide the EL method to improve its accuracy over that of batch EL. It also requires good strategies that suit both the EL task and the given dataset [11, 13].

Objective. In this paper, we introduce **interactive entity linking** to improve the embedding of entities which then yields better overall linking accuracy. In our proposed framework, we collect *useful* answers from human subjects as early as possible thereby reducing human efforts when they are limited. Since our EL is based on a pre-trained entity embedding model, it is interesting to see whether simultaneously updating the embedding and the EL method can improve accuracy. Our goal is to have the EL system interacts with annotators who contribute labels to enhance both the embedding model and EL method. To optimize the use of human efforts and minimize the labeling costs, we adopt an active learning (AL) approach. We propose two AL strategies and a hybrid strategy to select mentions for human annotation that can best improve EL accuracy through updating embedding model.

The following summarizes our research contributions: (1) We propose a framework that leverages on embedding retrofitting and human knowledge to improve current EL method. (2) We propose two mention selection strategies: Strategy 1 selects mentions that improve the coherence of results with other mentions, and Strategy 2 selects mentions that are difficult for embedding-based EL. We also derive Hybrid methods based on the two strategies. (3) We evaluate our interactive EL methods on both a benchmark dataset and a real-world dataset. The evaluation results show that our active learning strategies outperform random task selection.

2 Related Works

While there have been work on interactive entity linking, it focuses on interactive UI design for user to correct the EL result interactively [1]. This is very different from our proposed idea of directing users to correct only small number of selected EL results so as to improve the entity embeddings for linking other entity mentions in the input text more accurately. Active learning is a general learning strategy that performs instance selection to reduce the annotation cost [6]. The key idea of active learning is to iteratively assign unlabeled instances to a user/oracle for annotation to maximize the accuracy of a classifier with minimum annotation cost. Active learning has been widely used on crowdsourcing platforms in various supervised learning tasks such as classification and NER [11, 13, 15]. One important instance selection principle in active learning is **uncertainty sampling** [5] , that is, to select the most uncertain instance according to the current model to be annotated. The classifier can then be tuned to more accurately predict labels for the ambiguous data instances. Another

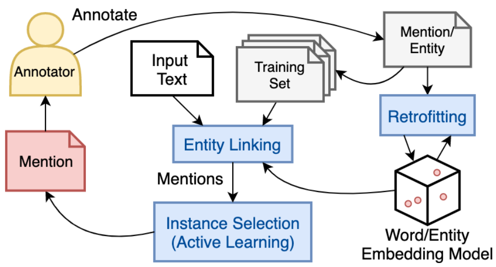

Figure 1: Proposed Interactive Entity Linking Framework

principle is **minimization of estimate of expected future error** [9]. An algorithm based on this principle selects for human annotation the instance that optimizes performance based on a perceived analytic expression of the expected error. Finally, **query by committee** [10] selects the instance which see disagreement among multiple independently trained models.

3 Interactive Entity Linking

Our proposed iterative EL approach includes two key ideas, namely: (a) retrofitting the entity-word embedding model to efficiently update the pre-trained embeddings with annotation labels, and (b) active learning to select mentions and their corresponding candidate entities that can contribute most to the overall accuracy. Figure 1 shows our proposed framework. In our framework, the **embedding learning** step requires an initial entity-word embedding model to be first learned for EL. In this work, we train the embedding model using Wikipedia2vec on Wikipedia corpus and further counterfit the model with synonym and antonym semantics [7, 12]. We call this new embedding model Wiki2Vec++. Using the learned entity-word embeddings, the **entity linking** step trains another model to link mentions in an input text to a dictionary of entities. The candidate entities for each mention are determined, and a confidence score is assigned to each mention-candidate entity pair. Here, we adopt Wiki2Vec++ embedding based EL method in [12] in two steps: (1) candidate generation that queries an entity-surface dictionary built from Wikipedia dump to generate a list of candidate entities and (2) training of Gradient Boosted regression tree ranker that generates confidence score for each candidate. For batch-mode EL, entity linking would stop here. For interactive EL, our framework continues for several steps to generate tasks for human annotation. The **instance selection** step creates a annotation task by selecting a mention (as instance) and its candidate entities. In this step, we adopt an active learning approach to determine the best mention and its candidate entities. As uncertainty in EL can be reduced by improving *linkage coherence* and *contextual closeness*, we propose instance selection strategies based on two proposed criteria to be described in Section 3.1.

Once the annotator completes an EL task, the **retrofitting** step is performed on the entity-word embedding model using entity-entity and entity-word semantic pairs extracted from human annotation as constraints. We will elaborate the retrofitting step in Section 3.2. With an updated embedding model, we repeat the earlier steps, including retraining of the EL method, to revise the EL results for the remaining mentions, and to identify the next mention(s) for annotation. The iterative process will terminate based on the following criteria: (1) the budget runs out (e.g., human annotator can only perform a fixed number of tasks); or (2) the EL accuracy is

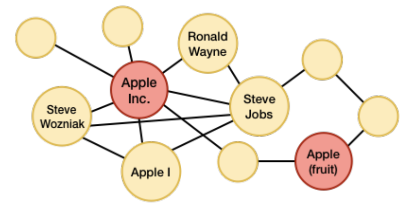

Figure 2: Entity Network of the "Apple" mention

good enough which can be measured by the consistency between human annotations and EL results.

3.1 Instance Selection

The goal of instance selection is to get the annotator to work on task instances that offer the highest utility to the EL method. Formally, we denote the set of all mentions, the set of mentions not yet annotated by human and the set of mentions already annotated by human as M, M^U and M^L respectively ($M = M^U \cup M^L$). Let E denote the set of all entities, and E_i ($\subseteq E$) denote the current ranking of candidate entities for mention $m_i \in M$. We want to find $m' \in M^U$ such that the annotation of m' by human will maximize the rank improvement of ground truth entities for the remaining mentions. The rank improvement measure is defined by: $\sum_{m_i \in M^U - \{m'\}} r_i - r_i'$ where r_i is the rank of ground truth entity of mention m_i in E_i, and r_i' refers to the rank of the ground truth entity after the EL method has been revised with the annotation of m'. As ground truth entities are not known for the mentions yet to be linked M^U, we adopt an active learning approach to reduce the uncertainty of EL. In this approach, we aim to reduce uncertainty of linking mentions in M^U. In the following, we propose two basic strategies S1 and S2, and a hybrid strategy. In each strategy, we devise a **uncertainty score** C_i for each mention in M^U so as to determine the most uncertain mention for annotation.

Basic Strategy 1 (S1): Improvement in Entity Coherence. Strategy S1 seeks to reduce the uncertainty in EL measured by the inconsistency between the ranking of candidate entities the EL method generates for a given mention and the ranking of candidate entities based on other mentions in the input text. This is also known as the *entity coherence* requirement of EL. Consider the knowledge graph example in Figure 2. When the input text has mentions already or likely linked to Steve Jobs and Ronald Wayne, it is more likely that an "apple" mention refers to Apple Inc. rather than Apple (fruit).

We measure **entity coherence** by the correlation between the EL rank order and the coherence rank order of candidate entities according to other already or likely linked mentions. Suppose the current candidate entities for a target mention m_i is $E_i = (e_{i1}, e_{i2}, \cdots, e_{i\ell})$. We assign each candidate entity e_{ij} a coherence score S_{ij} based on how coherent it is with the entities of already linked mentions M^L and the top ranked candidates of not-yet-linked mentions M^U. A high S_{ij} implies that the candidate entity e_{ij} has high coherence with other mentions. Suppose R_i denote the rank order $(1, 2, \cdots, \ell)$ of the candidate entities for mention $m_i \in M^U$ returned by the EL method, and R_i^{S1} denote rank order of candidates e_{ij}'s based on decreasing coherence score S_{ij}. The uncertainty score of m_i, C_i^{S1}, is thus defined as the inverse correlation between the EL rank order and entity coherence rank order: $C_i^{S1} = -Corr(R_i, R_i^{S1})$ where $Corr$ is the rank correlation. S1 therefore selects the mention m_{i^*} that has the largest uncertainty score $C_{i^*}^{S1}$ denoted by $m_{i^*} = argmax_{m_i \in M^U} C_i^{S1}$.

To compute S_{ij} for each candidate entity e_{ij}, we conduct a **weighted random walk** on a **weighted entity network** G_i. G_i consists of entities already linked to mentions in M^L and top candidate entities for mentions in M^U. We denote this set of entities by E_i^M. An edge (e_j, e_k) is created between candidate entities e_j and e_k in E_i^M if they co-occur in some Wikipedia article. An edge weight $wt_{j,k}$ is then obtained by the min-max normalized co-occurrence:

$$\mathrm{wt}_{j,k} = (n_{j,k} - \min_{e_{j'}, e_{k'} \in E} n_{j',k'})/(\max_{e_{j'}, e_{k'} \in E} n_{j',k'} - \min_{e_{j'}, e_{k'} \in E} n_{j',k'})$$

where $n_{p,q}$ is the co-occurrence count of e_p and e_q.

Given the weighted entity network G_i of each mention m_i, we perform η random walks from its top candidate entity e_{ij} to entities in $E_i^M - E_i$ based on the a transition probability defined as $P(e_q|e_p) = \frac{wt_{p,q}}{\sum_{e_r \in E_i} wt_{p,r}}$. With the random walks, each candidate entity $e_k \in E_i^M - E_i$ is then assigned the average walk length from e_{ij}, $L(e_{ij}, e_k) = Avg_{p_{(e_{ij},e_k)}} L(p_{(e_{ij},e_k)})$ where $L(p_{(e_{ij},e_k)})$ denotes the length of a walk from e_{ij} to e_k denoted by $p_{(e_{ij},e_k)})$. Each candidate entity is traversed for $\eta = 10$ times. We finally compute the coherence score of e_{ij} as $S_{ij} = Avg_{e_k \in E_i^M - E_i} L(e_{ij}, e_k)$ and obtain the rank order R_i^{S1} accordingly.

Basic Strategy 2 (S2): Selection of Difficult Mentions by Contextual Closeness. S2 aims to reduce uncertainty of EL attributed to mismatch of candidate entities and the mention's context, suggesting that the mention is difficult for the current embedding-based EL method. S2 therefore gets the annotator to annotate such kind of mentions to minimize the expected error in the future iteration. Let a mention m_i's top K ranked candidate entities be $E_i(K)$. S2 measures m_i's uncertainty by the average distance from its vector (or context representation) v_i^m to the top-K ranked candidate entities. If the top K candidate entities are not close to m_i, we can assume that the EL method does not give an accurate prediction, and m_i thus needs to be given to the annotator as a task. Hence, we adopt the loss function of retrofitting [3], and rank all mention-entity pairs by $C_i^{S2} = \sum_{e_{ij} \in E_i(K)} \left\| \mathbf{v}_i^m - \mathbf{v}_{ij}^e \right\|$ where $\mathbf{v}_i^m$ is the averaged vector of words/entities in the context window of m_i, and $\mathbf{v}_{ij}^e$ is the entity embedding of e_{ij}. Empirically, we set both window size and K to 5, and will explore other settings in our future work.

Hybrid Strategy: Weighted Strategy. This strategy combines S1 and S2 by returning mentions aggregated from S1 and S2 by defining a combined uncertainty score C_i^{Hy} for each mention m_i to be a weighted sum between the normalized uncertainty scores assigned to m_i by S1 and S2, denoted by $C_i^{Hy} = \alpha Norm(C_i^{S1}) + (1 - \alpha) Norm(C_i^{S2})$ where $0 \le \alpha \le 1$. When $\alpha = 0.5$, the hybrid method gives equal weights to S1 and S2. The normalized uncertainly score of mention m_i under S1 is defined by $Norm(C_i^{S1}) = 1 - (C_i^{S1} - \min_{i'} C_{i'}^{S1})/(\max_{i'} C_{i'}^{S1} - \min_{i'} C_{i'}^{S1})$ The normalized uncertainty score of mention m_i under S2 is defined in a similar way. We then assign the mention with the highest C_i^{Hy} value to the annotator.

3.2 Embedding Updating using Retrofitting

Once a mention is selected, we assign it as a task for annotation. The task includes the sentence containing the mention and its candidate entities. The annotator is asked to identify the correct entity for the mention. Here, we assume that the human annotator is an expert capable of making ground truth decisions.

With the annotation result, we update the entity-word embedding of the knowledge base to adapt it to the current EL task. Firstly, we extract synonym constraints from the annotations for the later retrofitting process. For example, when linking the mention "Apple" in the sentence (Note: mentions are underlined): "Apple was founded by Steve Jobs, Steve Wozniak in April 1976." Suppose the human annotator decides that Apple Inc. is the correct entity for the mention "Apple", and Steve Jobs is known to be linked to the mention "Steve Jobs". We will extract *entity-to-entity pairs*, e.g., (Apple Inc.,Steve Jobs) and *entity-to-word pairs*, e.g., (Apple Inc., "founded") from the earlier example sentence. Finally, the entity-word embedding model is then retrofitted with the extracted entity-entity/entity-word pairs. Retrofitting methods tune an embedding model based on some input synonymous(antonymous) embedding pairs, called constraints, to let the embeddings be closer to(further away from) each other [3, 7]. In this paper, we use **explicit retrofitting**, a retrofitting variant [4], to update not only embeddings involved in the constraints, but also their neighbors. For example, from the thesaurus we learn that "stupid" is an antonym to "smart". If "intelligent" is close to "smart" in the vector space, intuitively we know that "intelligent" and "stupid" should also be antonymous. In addition to antonym relation, we also know "intelligent" to be semantically similar to the synonyms of "smart". Explicit retrofitting therefore tunes word embeddings based on such knowledge.

4 Experiment

In our experiment, we use the following datasets: **(a) ACE Arabic News Dataset:** We selected a subset of Arabic related articles from the ACE news dataset [8] which consist of 101 mentions which are linked to ground truth Wikipedia entries. **(b) Wikipedia Dataset:** This dataset consists of the main content of a Wikipedia page about **Iraq War**[1] and it covers 1341 mentions. This page is excluded from the learning of entity-word embedding model and counter-fitted the embedding model to incorporate word and entity semantics [7]. In this article, all mentions and their ground truth linked entities are given. We keep 30% of the linked mentions as prior knowledge, and the remaining 70% mentions as the testing set.

Our experiments cover the following methods, namely: (1) **Strategy S1**, (2) **Strategy S2**, (3) **Hybrid Strategy with** $\alpha = 0.5$ which gives equal weight to Strategies S1 and S2, (4) **Hybrid Strategy with** $\alpha = 0.6$ which has ideal α setting determined by tuning, and (5) **Random Strategy**. All the above interactive EL methods include the same EL step with retraining. Our evaluation focuses on (1) how much improvement on the performance does each strategy contribute, and (2) the effect caused by retrofitting by annotation. In each iteration, we measure the accuracy using Micro-Accuracy for mentions in the test set that have not yet been annotated. Suppose $e_1, e_2, \cdots, e_{|M^U|}$ denote the predicted entities for not-yet-annotated mentions $m_1, m_2, \cdots, m_{|M^U|}$ respectively. The predicted entity e_i

[1] https://en.wikipedia.org/wiki/Iraq_War

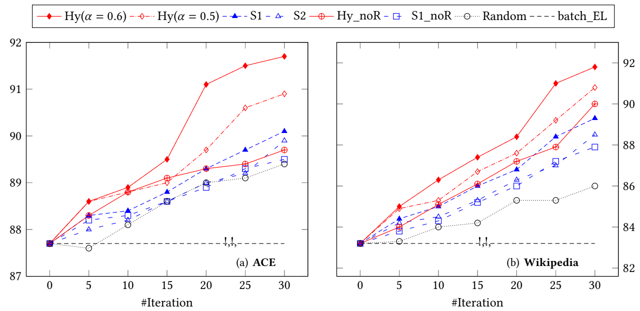

Figure 3: Entity Linking Results (Micro-Accuracy)

of mention m_i is the highest candidate entity ranked by the method. That is, Micro-Accuracy = $\frac{\#(m_i,e_i)\ pairs\ correctly\ linked}{|M^U|}$.

5 Results and Discussion

Figure 3 shows the accuracy results of the different EL methods from the 0^{th} iteration to the 30^{th} iteration. The x-axis represents the number of iterations (or number of mentions annotated). At the x^{th} iteration, the accuracy is measured after the x^{th} selected mention has been annotated and used to retrofit the entity-word embedding model. The grey horizontal dashed lines **batch_EL** (with micro-accuracy 87.7 and 83.2 for ACE and Wikipedia datasets respectively) shows the accuracy of the batch EL method without human annotation. This is also the same as that of interactive EL at the 0th iteration. As shown in Figure 3, all the methods (including Random) improve accuracy as more mentions are annotated. Across the four datasets, the hybrid methods **Hy(α =*)** outperform the single strategy methods **S1** and **S2**. Hybrid(α = 0.6) yields the best accuracy in all iterations. Between the single strategy methods, Strategy 1 consistently yields better accuracy than Strategy 2. This explains why giving more weight (α = 0.6) to Strategy 1 works better for Hybrid. The above results show that active learning-based methods choosing uncertain mentions for human annotation consistently outperform random selection by about 4% even when the base performance is already quite accurate (> 80%).

Interestingly, we find that poorly selected mentions can cause a toll on performance. For example, Random method observes a performance drop in the 0^{th} and 5^{th} iterations as shown in Figure 3a. Therefore, even after the annotator has contributed correctly linked mentions might not contribute to better result in the next iteration as the embedding features are affected in the retrofitting. One example of effective instance selection is in the Wikipedia dataset with input text "Financial costs with approximately \$612 billion spent as of 4/09 the CBO has estimated the total cost of the war in Iraq to the United States will be around \$1.9 trillion". Both mentions "Financial costs" and "CBO" are assigned high uncertainty scores from the initial EL classifier. In the first iteration, "CBO" is selected by the mention selection strategy and assigned to the worker. The annotator linked it to Congressional Budget Office, which ranks behind Arab League boycott of Israel in the initial EL result. After the embeddings and model are updated, we are able to correctly link "Financial costs" to Financial cost of the Iraq War instead of the general Cost as the former has a better linkage with Congressional Budget Office.

In addition, we investigate whether entity-word embedding updating using retrofitting helps to link entity more accurately. We modify the methods such that the entity-word embedding is not retrofitted after receiving annotated mentions from the human expert. In particular, we experiment the versions of Hybrid (α = 0.6) (**Hy_noR**) and Strategy 1 (**S1_noR**) based on the same static embedding model. From Figure 3 we can see that **Hy_noR** and **S1_noR** observe slower improvement in performance over the different iterations. The updating versions of Hybrid (α = 0.6) and Strategy 1 outperform the non-updating versions. The performance drop for Hybrid can be as large as 2% in ACE dataset in the 30^{th} iteration. In summary, updating of entity-word embedding generally contributes to improvement of accuracy.

6 Conclusion

In this paper, we propose an active learning approach to select mentions for human annotation. Our experiment results show that our proposed mention selection strategies, particularly the Hybrid ones, effectively finds the most informative mention to reduce uncertainty in subsequent entity linking iterations. We also conducted experiments on different settings to examine the effect of retrofitting. As part of future work, we will evaluate our embedding approach to EL using other embedding techniques. Finally, we will also consider EL under real crowd-sourcing scenario, where the workers have different expertise, and do not guarantee to provide correct answers. Thus, the stopping criteria and answer aggregation strategy will have to be further investigated.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings & conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Sebastian Arnold, Robert Dziuba, and Alexander Löser. 2016. Tasty: Interactive entity linking as-you-type. In *COLING*.
[2] Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs. In *ISWC*.
[3] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy, and N. A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *NAACL-HLT*.
[4] Goran Glavaš and Ivan Vulić. 2018. Explicit Retrofitting of Distributional Word Vectors. In *ACL*.
[5] David D Lewis and William A Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *SIGIR*.
[6] David JC MacKay. 1992. Information-based objective functions for active data selection. *Neural computation* 4, 4 (1992), 590–604.
[7] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting Word Vectors to Linguistic Constraints. In *NAACL-HLT*.
[8] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *HLT*.
[9] Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through monte carlo estimation of error reduction. In *ICML*.
[10] H S Seung, M Opper, and H Sompolinsky. 1992. Query by Committee. In *COLT*.
[11] Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems* 159 (2018), 244–258.
[12] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *CoNLL*.
[13] Y. Yan, R. Rosales, G. Fung, and J. G. Dy. 2011. Active learning from crowds. In *ICML*.
[14] S. W.-t. Yih, M.-W. Chang, X. He, and J. Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *ACL*.
[15] Jinhong Zhong, Ke Tang, and Zhi-Hua Zhou. 2015. Active Learning from Crowds with Unsure Option. In *IJCAI*.