Relation Extraction from Community Generated Question-Answer Pairs

Denis Savenkov
Emory University
dsavenk@emory.edu

Wei-Lwun Lu
Google
weilwunlu@google.com

Jeff Dalton
Google
jeffdalton@google.com

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Abstract

Community question answering (CQA) websites contain millions of question and answer (QnA) pairs that represent real users' interests. Traditional methods for relation extraction from natural language text operate over individual sentences. However answer text is sometimes hard to understand without knowing the question, *e.g.*, it may not name the subject or relation of the question. This work presents a novel model for relation extraction from CQA data, which uses discourse of QnA pairs to predict relations between entities mentioned in question and answer sentences. Experiments on 2 publicly available datasets demonstrate that the model can extract from $\sim$20% to $\sim$40% additional relation triples, not extracted by existing sentence-based models.

1 Introduction

Recently all major search companies have adopted knowledge bases (KB), and as a result users now can get rich structured data as answers to some of their questions. However, even the largest existing knowledge bases, such as Freebase (Bollacker et al., 2008), DPpedia (Auer et al., 2007), NELL (Carlson et al., 2010), Google Knowledge Graph *etc.*, which store billions of facts about millions of entities, are far from being complete (Dong et al., 2014). A lot of information is hidden in unstructured data, such as natural language text, and extracting this information for knowledge base population (KBP) is an active area of research (Surdeanu and Ji, 2014).

One particularly interesting source of unstructured text data is CQA websites (*e.g.* Yahoo! Answers,[1] Answers.com,[2] *etc.*), which became very popular resources for question answering. The information expressed there can be very useful, for example, to answer future questions (Shtok et al., 2012), which makes it attractive for knowledge base population. Although some of the facts mentioned in QnA pairs can also be found in some other text documents, another part might be unique (*e.g.* in Clueweb[3] about 10% of entity pairs with existing Freebase relations mentioned in Yahoo!Answers documents cannot be found in other documents). There are certain limitations in applying existing relation extraction algorithms to CQA data, *i.e.*, they typically consider sentences independently and ignore the discourse of QnA pair text. However, often it is impossible to understand the answer without knowing the question. For example, in many cases users simply give the answer to the question without stating it in a narrative sentence (*e.g.* "*What does "xoxo" stand for? Hugs and kisses.*"), in some other cases the answer contains a statement, but some important information is omitted (*e.g.* "*What's the capital city of Bolivia? Sucre is the legal capital, though the government sits in La Paz*").

In this work we propose a novel model for relation extraction from CQA data, that uses discourse of a QnA pair to extract facts between entities mentioned in question and entities mentioned in answer sentences. The conducted experiments confirm that many of such facts cannot be extracted by existing sentence-based techniques and thus it is beneficial to combine their outputs with the output of our model.

2 Problem

This work targets the problem of relation extraction from QnA data, which is a collection of (q, a) pairs,

[1] http://answers.yahoo.com/

[2] http://www.answers.com

[3] http://www.lemurproject.org/clueweb12/

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 96–102,
Denver, Colorado, June 1, 2015. ©2015 Association for Computational Linguistics

where q is a question text (can contain multiple sentences) and a is the corresponding answer text (can also contain multiple sentences). By relation instance r we mean an ordered binary relation between $subject$ and $object$ entities, which is commonly represented as $[subject, predicate, object]$ triple. For example, the fact that Brad Pitt married Angelina Jolie can be represented as [Brad Pitt, married_to, Angelina Jolie]. In this work we use Freebase, an open schema-based KB, where all entities and predicates come from the fixed alphabets E and P correspondingly. Let e_1 and e_2 be entities that are mentioned together in a text (*e.g.* in a sentence, or e_1 in a question and e_2 in the corresponding answer), we will call such an entity pair with the corresponding context a mention. The same pair of entities can be mentioned multiple times within the corpus, and for all mentions $i = 1, ..., n$ the goal is to predict the expressed predicate ($z_i \in P$) or to say that none applies ($z_i = \emptyset$). Individual mention predictions $z_1, ..., z_n$ are combined to infer a set of relations $\mathbf{y} = \{y_i \in P\}$ between the entities e_1 and e_2.

3 Models

Our models for relation extraction from QnA data incorporates the topic of the question and can be represented as a graphical model (Figure 1). Each mention of a pair of entities is represented with a set of mention-based features x and question-based features x_t. A multinomial latent variable z represents a relation (or none) expressed in the mention and depends on the features and a set of weights w_x for mention-based and w_t for question-based features: $\hat{z} = \underset{z \in P \cup \emptyset}{arg\,max}\; p(z|x, x_t, w_x, w_t)$. To estimate this variable we use L2-regularized multinomial logistic regression model, trained using the distant supervision approach for relation extraction (Mintz et al., 2009), in which mentions of entity pairs related in Freebase are treated as positive instances for the corresponding predicates, and negative examples are sampled from mentions of entity pairs which are not related by any of the predicates of interest. Finally, to predict a set of possible relations $\mathbf{y}$ between the pair of entities we take logical OR of individual mention variables $\mathbf{z}$, *i.e.* $y_p = \vee_{i=1}^{M}[z_i = p, p \in P]$, where M is the number of mentions of this pair of entities.

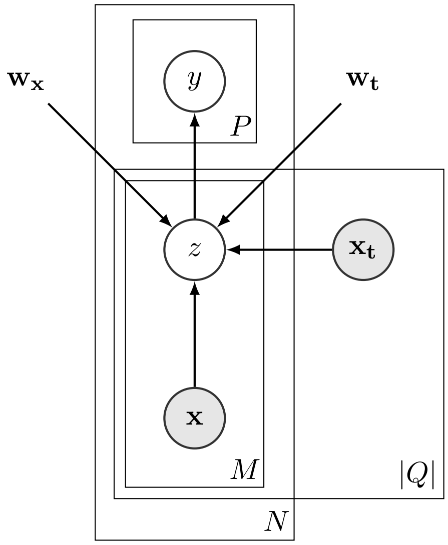

Figure 1: QnA-based relation extraction model plate diagram. N - number of different entity pairs, M - number of mentions of an entity pair, $|Q|$ - number of questions where an entity pair is mentioned, $\mathbf{x}$ and $\mathbf{x_t}$ - mention-based and question-based features, $\mathbf{w}$ and $\mathbf{w_t}$ - corresponding feature weights, latent variables z - relation expressed in an entity pair mention, latent variables y - relations between entity pair

3.1 Sentence-based baseline model

Existing sentence-based relation extraction models can be applied to individual sentences of a QnA pair and will work well for complete statements, *e.g.* “Who did Brad Pitt marry? Brad Pitt and Angelina Jolie married at secret ceremony”. In sentence-based scenario, when the set of question-based features is empty, the above model corresponds to the Mintz++ baseline described in Surdeanu et al. (2012), which was shown to be superior to the original model of Mintz et al. (2009), is easier to train than some other state of the art distant supervision models and produces comparable results.

3.2 Sentence-based model with question features

In many cases an answer statement is hard to interpret correctly without knowing the corresponding question. To give the baseline model some knowledge about the question, we include question features (Table 1), which are based on dependency tree and surface patterns of a question sentence. This

Table 1: Examples of features used for relation extraction for "*When was Mariah Carey born? Mariah Carey was born 27 March 1970*"

Sentence-based model	
Dependency path between entities	[PERSON]→nsubjpass(born)tmod←[DATE]
Surface pattern	[PERSON] be/VBD born/VBN [DATE]

Question features for sentence-based model	
Question template	when [PERSON] born
Dependecy path from a verb to the question word	(when)→advmod(born)
Question word + dependency tree root	when+born

QnA-based model	
Question template + answer entity type	Q: when [PERSON] born A:[DATE]
Dependency path from question word to entity and answer entity to the answer tree root	Q:(when)→advmod(born)nsubj←[PERSON] A: (born)tmod←[DATE]
Question word, dependency root and answer pattern	Q: when+born A:born [DATE]

information can help the model to account for the question topic and improve predictions in some ambiguous situations.

3.3 QnA-based model

The QnA model for relation extraction is inspired by the observation, that often an answer sentence do not mention one of the entities at all, *e.g.*, "*When was Isaac Newton born? December 25, 1642 Woolsthorpe, England*". To tackle this situation we make the following assumption about the discourse of a QnA pair: an entity mentioned in a question is related to entities in the corresponding answer and the context of both mentions can be used to infer the relation predicate. Our QnA-based relation extraction model takes an entity from a question sentence and entity from the answer as a candidate relation mention, represents it with a set features (Table 1) and predicts a possible relation between them similar to sentence-based models. The features are conjunctions of various dependency tree and surface patterns of question and answer sentences, designed to capture their topics and relation.

4 Experiments

4.1 Datasets

For experiments we used 2 publicly available CQA datasets: Yahoo! Answers Comprehensive Questions and Answers[4] and a crawl of WikiAnswers[5] (Fader et al., 2014). The Yahoo! Answers dataset contains 4,483,032 questions (3,894,644 in English) with the corresponding answers collected on 10/25/2007. The crawl of WikiAnswers has 30,370,994 question clusters, tagged by WikiAnswers users as paraphrases, and only 3,386,256 them have answers. From these clusters we used all possible pairs of questions and answers (19,629,443 pairs in total).

For each QnA pair we applied tokenization, sentence detection, named entity tagger, parsing and coreference resolution from Stanford CoreNLP (Manning et al., 2014). Our cascade entity linking approach is similar to Chang et al. (2011) and considered all noun phrase and named entity mentions as candidates. First all named entity mentions are looked up in Freebase names and aliases dictionary. The next two stages attempt to match mention text with dictionary of English Wikipedia concepts (Spitkovsky and Chang, 2012) and its normalized version. Finally for named entity mentions we try spelling correction using Freebase entity names dictionary. We didn't disambiguate entities and instead took top-5 ids for each coreference cluster (using the $p(entity|phrase)$ score from the dictionary or number of existing Freebase triples). All pairs of entities (or entity and date) in a QnA pair that are directly related[6] in Freebase were annotated with the corresponding relations.

[4] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[5] http://wiki.answers.com

[6] We also consider some paths that come through a mediator node, *e.g.*/people/person/spouse_s./people/marriage/spouse

Table 2: Yahoo! Answers and WikiAnswers datasets statistics

	Y!A	WA
Number of QnA pairs	3.8M	19.6M
Average question length (in chars)	56.67	47.03
Average answer length (in chars)	335.82	24.24
Percent of QnA pairs with answers that do not have any verbs	8.8%	18.9%
Percent of QnA pairs with at least one pair of entities related in Freebase	11.7%	27.5%
Percent of relations between entity pairs in question sentences only	1.6 %	3.1%
Percent of relations between entity pairs in question and answer sentences only	28.1%	46.4%
Percent of relations between entity pairs in answer sentences only	38.6%	12.0%

Table 2 gives some statistics on the datasets used in this work. The analysis of answers that do not have any verbs show that ∼8.8% of all QnA pairs do not state the predicate in the answer text. The percentage is higher for WikiAnswers, which has shorter answers on average. Unfortunately, for many QnA pairs we were unable to find relations between the mentioned entities (for many of them no or few entities were resolved to Freebase). Among those QnA pairs, where some relation was annotated, we looked at the location of related entities. In Yahoo! Answers dataset 38.6% (12.0% for WikiAnswers) of related entities are mentioned in answer sentences and can potentially be extracted by sentence-based model, and 28.1% (46.4% for WikiAnswers) between entities mentioned in question and answer sentences, which are not available to the baseline model and our goal is to extract some of them.

4.2 Experimental setup

For our experiments we use a subset of 29 Freebase predicates that have enough unique instances annotated in our corpus, *e.g.* date of birth, profession, nationality, education institution, date of death, disease symptoms and treatments, book author, artist album, *etc.* We train and test the models on each dataset separately. Each corpus is randomly split for training (75%) and testing (25%). Knowledge base facts are also split into training and testing sets (50% each). QnA and sentence-based models predict labels for each entity pair mention, and we aggregate mention predictions by taking the maximum score for each predicate. We do the same aggregation to produce a combination of QnA- and sentence-based models, *i.e.*, all extractions produced by the models are combined and if there are multiple extractions of the same fact we take the maximum score as the final confidence. The precision and recall of extractions are evaluated on a test set of Freebase triples, *i.e.* an extracted triple is considered correct if it belongs to the test set of Freebase triples, which are not used for training (triples used for training are simply ignored). Note, that this only provides a lower bound on the model performance as some of the predicted facts can be correct and simply missing in Freebase.

4.3 Results

Figure 2 shows Precision-Recall curves for QnA-based and sentence-based baseline models and some numeric results are given in Table 3. As 100% recall we took all pairs of entities that can be extracted by either model. It is important to note, that since some entity pairs occur exclusively inside the answer sentences and some in pairs of question and answer sentences, none of the individual models is capable of achieving 100% recall, and maximum possible recalls for QnA- and sentence-based models are different.

Results demonstrate that from 20.5% to 39.4% of correct triples extracted by the QnA-based model are not extracted by the baseline model, and the combination of both models is able to achieve higher precision and recall. Unfortunately, comparison of sentence-based model with and without question-based features (Figure 2) didn't show a significant difference.

5 Error analysis and future work

To get an idea of typical problems of QnA-based model we sampled and manually judged extracted high confidence examples that are not present in

Table 3: Extraction results for QnA- and sentence-based models on both datasets

	Yahoo! Answers			WikiAnswers		
	QnA	Sentence	Combined	QnA	Sentence	Combined
F-1 score	0.219	0.276	0.310	0.277	0.297	0.332
Number of correct extractions	3229	5900	7428	2804	2288	3779
Correct triples not extracted by other model	20.5%	56.5%	-	39.4%	25.8%	-

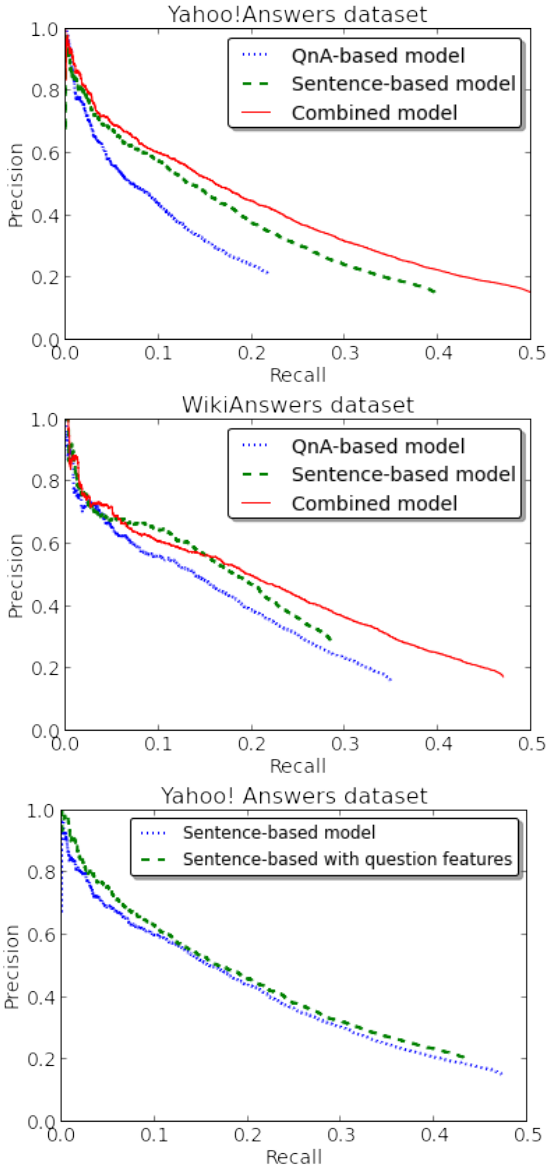

Figure 2: Precision-Recall curves for QnA-based vs sentence-based models and sentence-based model with and without question features

Freebase (and thus are considered incorrect for precision-recall analysis).

The major reason (40%) of false positive extractions is errors in entity linking. For example: "*Who is Tim O'Brien? He was born in Austin on October 1, 1946*". The model was able to correctly extract [Tim O'Brien, date_of_birth, October 1, 1946], however Tim O'Brien was linked to a wrong person. In a number of cases (16%) our discourse model turns out to be too simple and fails for answers, that mention numerous additional information, *e.g.* "*How old is Madonna really? ...Cher was born on 20 May 1946 which makes her older that Madonna...*". A possible solution would be to either restrict QnA-based model to cases when no additional information is present or design a better discourse model with deeper analysis of the answer sentence and its predicates and arguments. Some mistakes are due to distant supervision errors, for example for the music.composition.composer predicate our model extracts singers as well as composers (which are in many cases the same).

Of course, there are a number of cases, when our extractions are indeed correct, but are either missing (33%) or contradicting with Freebase (8%). An example of an extracted fact, that is missing in Freebase is "*Who is Wole Soyinka? He studied at the University College, Ibadan(1952-1954) and the University of Leeds (1954-1957)*", and [Wole Soyinka, institution, University of Leeds] is currently not present in Freebase. Contradictions with Freebase occur because of different precision levels ("pianist" vs "jazz pianist", city vs county, *etc.*), different calendars used for dates or "incorrect" information provided by the user. An example, when existing and extracted relation instance are different in precision is:"*Who is Edward Van Vleck? Edward Van Vleck was a mathematician born in Middletown, Connecticut*" we extract [Edward Van Vleck, place_of_birth, Middletown], however the Freebase currently has USA as his place of birth.

The problem of "incorrect" information provided in the answer is very interesting and worth special attention. It has been studied in CQA research, *e.g.* (Shah and Pomerantz, 2010), and an example of such QnA pair is: "*Who is Chandrababu Naidu? Nara Chandra Babu Naidu (born April 20, 1951)*". Other authoritative resources on the Web give April 20, 1950 as Chandrababu Naidu's date of birth. This raises a question of trust to the provided answer and expertise of the answerer. Many questions on CQA websites belong to the medical domain, *e.g.* people asking advices on different health related topics. How much we can trust the answers provided to extract them into the knowledge base? We leave this question to the future work.

Finally, we have seen that only a small fraction of available QnA pairs were annotated with existing Freebase relations, which shows a possible limitation of Freebase schema. A promising direction for future work is automatic extraction of new predicates, which users are interested in and which can be useful to answer more future questions.

6 Related work

Relation extraction from natural language text has been an active area of research for many years, and a number of supervised (Snow et al., 2004), semi-supervised (Agichtein and Gravano, 2000) and unsupervised (Fader et al., 2011) methods have been proposed. These techniques analyze individual sentences and can extract facts stated in them using syntactic patterns, sentence similarity, *etc*. This work focus on one particular type of text data, *i.e.* QnA pairs, and the proposed algorithm is designed to extract relations between entities mentioned in question and answer sentences.

Community question-answering data has been a subject of active research during the last decade. Bian et al. (2008) and Shtok et al. (2012) show how such data can be used for question answering, an area with a long history of research, and numerous different approaches proposed over the decades (Kolomiyets and Moens, 2011). One particular way to answer questions is to utilize structured KBs and perform semantic parsing of questions to transform natural language questions into KB queries. Berant et al. (2013) proposed a semantic parsing model that can be trained from QnA pairs, which are much easier to obtain than correct KB queries used previously. However, unlike our approach, which takes noisy answer text provided by a CQA website user, the work of Berant et al. (2013) uses manually created answers in a form of single or lists of KB entities. Later Yao and Van Durme (2014) presented an information extraction inspired approach, that predicts which of the entities related to an entity in the question could be the answer to the question. The key difference of this work from question answering is that our relation extraction model doesn't target question understanding problem and doesn't necessarily extract the answer to the question, but rather some knowledge it can infer from a QnA pair. Many questions on CQA websites are not factoid, and there are many advice and opinion questions, which simply cannot be answered with a KB entity or a list of entities. However, it is still possible to learn some information from them (*e.g.* from "*What's your favorite Stephen King book? The Dark Half is a pretty incredible book*" we can learn that the Dark Half is a book by Stephen King). In addition, answers provided by CQA users often contain extra information, which can also be useful (*e.g.* from "*Where was Babe Ruth born? He was born in Baltimore, Maryland on February 6th, 1895*" we can learn not only place of birth, but also date of birth of Babe Ruth).

7 Conclusion

In this paper we proposed a model for relation extraction from QnA data, which is capable of predicting relations between entities mentioned in question and answer sentences. We conducted experiments on 2 publicly available CQA datasets and showed that our model can extract triples not available to existing sentence-based techniques and can be effectively combined with them for better coverage of a knowledge base population system.

Acknowledgments

This work was funded by the Google Faculty Research Award. We gratefully thank Evgeniy Gabrilovich and Amar Subramanya for numerous valuable and insightful discussions, and anonymous reviewers for useful comments on the work.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, pages 1533–1544.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 467–476, New York, NY, USA. ACM.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr., and T.M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, AAAI'10, pages 1306–1313. AAAI Press.

Angel X Chang, Valentin I Spitkovsky, Eneko Agirre, and Christopher D Manning. 2011. Stanford-ubc entity linking at tac-kbp, again. In *Proceedings of Text Analysis Conference*, TAC'11.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA. ACM.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1156–1165, New York, NY, USA. ACM.

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Inf. Sci.*, 181(24):5412–5434, December.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09.

Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 411–418. ACM.

Anna Shtok, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. Learning from the past: Answering new questions with past answers. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 759–768, New York, NY, USA. ACM.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Valentin I Spitkovsky and Angel X Chang. 2012. A cross-lingual dictionary for english wikipedia concepts. In *LREC*, pages 3168–3175.

Mihai Surdeanu and Heng Ji. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. Text Analysis Conference (TAC2014)*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 455–465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of ACL*, ACL'14.