Using Natural Language to Integrate, Evaluate, and Optimize Extracted Knowledge Bases

Doug Downey, Chandra Sekhar Bhagavatula
Northwestern University, Evanston, IL
ddowney@eecs.northwestern.edu,
csbhagav@u.northwestern.edu

Alexander Yates
Temple University, Philadelphia, PA
alexander.yates@temple.edu

ABSTRACT

Web Information Extraction (WIE) systems extract billions of unique facts, but integrating the assertions into a coherent knowledge base and evaluating across different WIE techniques remains a challenge. We propose a framework that utilizes natural language to integrate and evaluate extracted knowledge bases (KBs). In the framework, KBs are integrated by exchanging probability distributions over natural language, and evaluated by how well the output distributions predict held-out text. We describe the advantages of the approach, and detail remaining research challenges.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval; H.3.5 [**Information Storage and Retrieval**]: Online Information ServicesWeb-based services

Keywords

Knowledge Extraction; Knowledge Integration; Language Modeling

1. INTRODUCTION

Extracting knowledge automatically from the Web is known as Web Information Extraction (WIE), and is a task of broad and increasing interest. Over the past decade, a variety of research studies and prototypes have investigated WIE techniques [1–11]. WIE has recently been pursued in industry in the form of question-answering systems like IBM's Watson [12] and Web search aids such as the Google Knowledge Graph and Microsoft Satori. WIE presents a promising route toward achieving Tim Berners-Lee's vision of a Semantic Web, and one day acquiring the knowledge needed to enable human-level artificial intelligence.

Existing WIE systems vary along two key dimensions: the *type of content* they target for extraction (Web tables, text, Wikipedia, etc.), and the *representation of the extracted knowledge* (individual tuples or frames, or additions to given ontology). Because WIE systems are so diverse, it is difficult to integrate knowledge across extracted KBs: as discussed below, existing schema matching techniques appear to be insufficient (see Section 3). Further, it is unclear how to *evaluate* across extraction approaches that target incomparable knowledge representation schemes. To deliver on the promise of WIE, new methods are required that allow different system builders to work together to construct a massive body of knowledge.

In this position paper, we propose a framework for integrating and evaluating WIE systems. The approach hinges on representing extracted knowledge in terms of *probability distributions over natural language (NL)*. Many existing WIE systems already utilize such distributions as input, at least implicitly—as a simple example, the distribution of terms C and x in the extraction pattern "C such as x" is commonly used to extract x's that are members of the class C, as in the phrase "cities such as Boston" [2, 13]. Assertions that occur more frequently in text (i.e., for which the extraction pattern has higher probability) are deemed more likely to be correct [14]. Our contention in this paper is that a generalization of this capability, in which KBs import and export distributions over language, can enable automated integration of WIE systems. Further, we believe that the quality of the output distributions (according to some measure) forms a promising metric for evaluating and optimizing WIE systems.

We envision a large-scale research effort in which different parties continuously extract KBs in a variety of ways, and the KBs selectively share knowledge with each other in natural language in an effort to encode a vast, high-precision, globally-interoperable body of knowledge. As discussed in Sections 3 and 4 below, utilizing NL for knowledge base integration and evaluation has distinct advantages: it enables KB integration without requiring a commitment to any single ontology, and it enables KB optimization over trillions of readily available evaluation examples (in the form of running text on the Web). However, operationalizing the proposed approach entails a number of research challenges, detailed in Section 5. We begin by discussing previous work in WIE.

2. WEB INFORMATION EXTRACTION

Web Information Extraction (WIE) is the task of extracting knowledge from content on the Web. Different WIE approaches target different types of Web content. Some systems extract knowledge from text across the Web [15], while others focus on Wikipedia content [16, 17] or Web tables

AKBC'13, October 27-28 2013, San Francisco, CA, USA

ACM 978-1-4503-2411-3/13/10 ...$15.00.
http://dx.doi.org/10.1145/2509558.2509569.

[5, 6, 18]. Other approaches integrate knowledge solicited from Web contributors [19]. Together, these KBs contain billions of facts, spanning an enormous variety of topics.

WIE approaches also differ significantly in the degree of representational structure in the extracted knowledge. For example, some approaches extract independent propositions or tuples (e.g. `MayorOf(Bloomberg, New York City)`) [4, 20], while others extract more comprehensive semantic frames [11]. Some approaches organize extracted facts into an ontology: these range from lightweight ontologies, often rooted in Wikipedia [7–10], to rich knowledge representation systems such as Cyc [1, 3].

The diversity of extracted facts and knowledge representation schemes presents a significant difficulty: it is unclear how to best combine different systems or evaluate across different systems. We present our proposed solution to these problems in the following sections.

3. NL FOR INTEGRATING KBS

If two KBs contain different sets of knowledge, represented in different ways, how can the KBs share knowledge with each other?

Previous work on this task includes *data integration* approaches from the database community, which attempt to merge two different KBs into a single knowledge base [21]. This approach is limited in that it generally requires special-purpose engineering or training examples for each *pair* of KBs to be integrated.

A distinct, potentially more scalable approach involves choosing one or more common reference ontologies with which all other KBs can be integrated. This approach is employed in the Linked Open Data project [22], in which different knowledge bases link their statements to a handful of common shared vocabularies. Wikipedia is a typical reference ontology for this task, and semi-automated methods for integrating with Wikipedia have been proposed for Cyc [23], relational databases [24], and tuples extracted from text [25]. This approach, while potentially much more scalable than pairwise integration, is also heavily restrictive: a small number of reference knowledge bases must be selected, and choosing such KBs is difficult. Further, changes to a reference knowledge base can entail burdensome updates to how each KB exports knowledge. While we believe integration with Linked Open Data is an important component of KB integration (and we discuss how to incorporate it in our approach in Section 5), we believe natural language integration has distinct advantages as discussed below.

3.1 The NL Protocol

We propose an approach in which knowledge bases are integrated by exchanging natural language. Because KBs will typically have uncertainty associated with their knowledge and how it is expressed in language, we define a protocol that exchanges not raw text but instead probability distributions over language.

We begin by giving a concrete example of how two knowledge bases can employ the protocol to exchange knowledge. We then provide a formal definition of the protocol, and discuss its advantages.

3.1.1 Concrete Example

Consider a knowledge base K with an objective of constructing a list of cities containing skyscrapers. Assume K extracts information from Wikipedia tables, and it has not found a table specifically listing this information. However, K also has a set of lexical extraction patterns (as in e.g. [2]), and utilizes these to determine it desires strings x that yield high values of the product:

$$P(x \text{ and other cities}) * P(\text{skyscraper in } x) \qquad (1)$$

Given reliable probability estimates of the above expression, K can estimate which x's are, in fact, cities with skyscrapers using existing WIE techniques [2, 14].

K lacks its own textual corpus, so it employs the NL protocol to pose a *query* to another KB K'. Specifically, K sends Equation 1 to K', and requests that K' return a list of strings x and the estimated value of the product for each.

Assume K' extracts from a textual corpus that includes three answers with positive probability: Shanghai, New York City, and Montreal. For concreteness, K' simply returns a distribution where each of these three strings has probability 0.2, and the remaining 0.4 of probability mass is distributed uniformly over other phrases.

Based on the three "seed" examples with non-negligible probability, the table extractor K can attempt to estimate a more accurate distribution over x using its KB of tables along with the *distributional hypothesis*, the notion that terms with similar meanings tend to appear in similar contexts [26]. Specifically, given a table column that contains the seed cities, the other cells S in the same column are likely to be semantically similar to the seeds. Thus, K can adjust its distribution over x to give higher probability to the strings S, and thereby estimate a more accurate distribution. K then runs an existing WIE technique to estimate, from the frequency information, which of the x's are correct assertions (that is, cities with skyscrapers). For simplicity, we make the reasonable assumption that the WIE technique concludes that the strings in S are cities with skyscrapers, and all other strings are not.

In the example, the first table returned by the WikiTables extraction system [27] when queried for the three seeds lists the top 40 cities ranked in terms of "Global City Competitiveness Index." While this table does not explicitly refer to skyscrapers, it happens that all 40 listed cities do, in fact, contain skyscrapers. Thus, through the use of the NL protocol coupled with distributional similarity, K is able to compute a list of cities with skyscrapers that has perfect precision and dramatically higher recall than the original three strings returned by K'.

This example illustrates how even when a knowledge base may not contain the target relation or be based on text (e.g. K need not contain any table listing all cities with skyscrapers), it can leverage other KBs (in this case K') through the NL protocol. By exchanging knowledge, the two KBs in the example are able to produce an answer to a query that is dramatically better than either of the KBs could produce in isolation.

3.2 NL Protocol: Formal Definition

Here, we present a candidate definition of the NL protocol, which is capable of expressing the queries described in the example above. A wide variety of other NL methods are also conceivable for integrating KBs, and we discuss potential variants of the protocol later in Section 5.

A KB implements the NL protocol by issuing and responding to queries. Formally, a *query* in the NL protocol is:

1. A set F of zero or more *variable symbols*, along with a numeric *length range* in tokens for each variable.
2. A product of language probabilities $\Pi_i P(\mathbf{w}_i)$, where each w_i is a sequence of tokens, and each token is either a single textual word or a variable symbol.
3. An integer N specifying the number of highest-probability results to return.

The *response* to an NL query consists of a series of N tuples, where each tuple contains $|F|$ strings (the string values of the variables in the query) and an estimate for the requested product of probabilities. Each of the strings is required to have a length in tokens within the range specified for its variable in the query, and the N tuples returned represent the estimated highest-probability substitutions for the variables in the given probability expression.

We explicitly choose to allow *products* of probability expressions (query component #2 above) because they enable much more compact query responses. Consider the query in Equation 1. The querying KB K could instead issue two separate queries, one for $P(x \text{ and other cities})$ and another for $P(\text{skyscraper in } x)$, and then multiply the results to obtain the same response as in the original query. However, in order to ensure the top R values for the product are in fact obtained in this way, K would need to request a large number of tuples $N >> R$ for both queries, whereas the original query requires only $N = R$.

3.3 Advantages of the NL protocol

The primary advantages of the NL protocol are threefold. First, the communication medium (natural language) is extremely **expressive**, and not tied to any single ontology. Each KB must implement two methods, for reading and writing knowledge in natural language, and it can then integrate with every other KB implementing the protocol. Secondly, knowledge exchanged in the NL protocol is **durable**: a fixed distribution $P_K(\mathbf{w})$ output by K at some point in time remains informative, even if K later changes radically. Third, the protocol is readily **interpretable** by humans, because it is expressed in natural language.

4. NL FOR EVALUATING AND OPTIMIZING KBS

How to estimate the quality of an extracted KB is an open question. Previous work has emphasized that extracted KBs must ultimately be evaluated in terms of "end-tasks," such as decision making and question answering [28]. While end task evaluation is necessary to ensure the knowledge can yield useful technology, it can be cost-prohibitive: evaluation with end-tasks generally requires direct human judgments of performance. Unless we solicit new human input often, we risk overfitting to a fixed objective. Thus, while we believe end task evaluations are vital to periodically evaluate competing approaches, due to their cost they cannot be used for continuous optimization of KB constructors.

4.1 The NL Objective

We propose the **NL Objective** for KBs, in which a KB K is evaluated in terms of how accurately its output distribution $P_K(\mathbf{w})$ predicts held-out sentences $\mathbf{w}$. This objective has the advantage of not being biased toward a limited data set; in fact, *trillions* of examples for training and evaluation are available, in the form of running text on the Web.

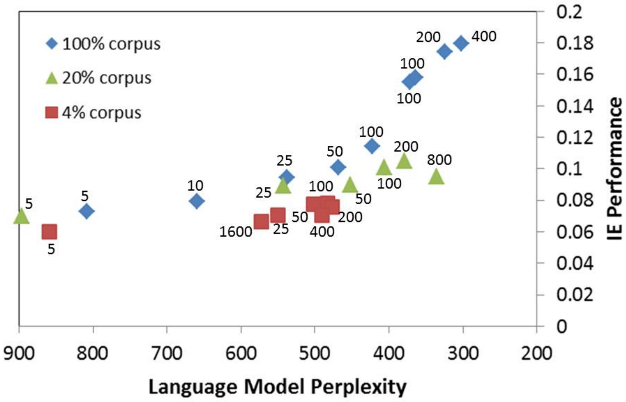

Figure 1: Web Information Extraction (WIE) performance of a Hidden Markov Model, as accuracy of the HMM's language model $P(\mathbf{w})$ varies (reprinted from [30]). Number labels indicate the number of latent states in the HMM, and performance is shown for three training corpus sizes (the full corpus consists of approximately 60 million tokens). WIE accuracy (in terms of area under the precision-recall curve) tends to increase as language modeling accuracy improves (i.e. perplexity decreases). WIE accuracy correlates more strongly with perplexity (-0.88, Spearman's) than with corpus size (0.71) or number of latent states (0.38).

A natural concern regarding the NL Objective is that what it rewards—the accuracy of language prediction—is only a surrogate for our ultimate goal of high-quality KBs. Predicting language can entail unproductive activities, such as mimicking "reporting bias" [29] (the NL Objective involves predicting what people *choose to say*, rather than what is true in the world) and predicting misinformation (e.g. estimating exactly how often people will assert that "Elvis killed JFK"[1]).

Despite the limitations, we believe the NL Objective has substantial utility for three reasons. First, while reporting bias and misinformation do exist, for a broad set of fact extraction mechanisms, veracity does *tend* to increase with the frequency of assertion [2, 14]. In fact, empirical evidence suggests that optimizing the standard perplexity metric in language modeling can lead to corresponding increases in WIE performance measures like precision and recall of extractions (see Figure 1). Secondly, it may be possible to mitigate misinformation by adding a component to the NL Objective that models the credibility of the Web site being predicted. Lastly, as discussed above, other evaluation schemes are arguably even more problematic. We discuss issues with the NL Objective further in Section 5.

5. KEY CHALLENGES

The NL Protocol and NL Objective present advantages as well as unique challenges. In this section, we detail key limitations of the techniques and remaining research challenges.

[1] This phrase occurs more than 30,000 times on the Web according to a major search engine.

5.1 Developing APIs in the NL Protocol

A primary limitation of the NL Protocol is that—like natural language, but unlike Linked Open Data RDF—messages in the NL Protocol are ambiguous. Given only the statement that "Chicago is a city," it is unclear whether the string Chicago refers to a fictional city or a real one, or whether it is the same meaning of Chicago in the phrase "that song by Chicago was playing on the radio."

We believe the ambiguity of language can be overcome to build a powerful protocol, through the use of well-designed queries. The query in the example in Section 3, for example, mitigates ambiguity by querying for a product of two distinct language probabilities. It is less likely that an erroneous or ambiguous phrase occurs in *both* the skyscraper and the city context, when compared to a single context alone. By composing larger products of additional indicative phrases (e.g. "cities including x," "x and other cities," etc.) we would expect that, as the number of phrases increases and if probability estimates are accurate, the high-probability x's would correspond to correct answers for any query. In fact, it can be proved that under assumptions that hold approximately in large corpora, such extraction techniques are *guaranteed* to achieve high accuracy [14].

To make sense disambiguation explicit, it would be possible to augment the NL protocol to allow not only surface strings, but also word classes of various types (e.g. parts of speech or well-established semantic classes). One option is to allow terms in queries that are not surface strings, but instead indicate a reference to a particular Linked Data URI, e.g., "x is the mayor of <reference to en.wikipedia.org/Chicago, Illinois>." This would allow KBs to leverage Linked Open Data URIs where they are well-established, but back off to natural language in other cases. Of course, utilizing URIs sacrifices some of the advantages of the NL protocol discussed above. New methods are needed to determine when utilizing URIs rather than "pure" NL is appropriate.

Lastly, while in the NL protocol we focus on single phrases $\mathbf{w}$, the API can be generalized to a richer discourse model. One simple example would allow distributions $P(\mathbf{w}|\mathbf{t})$, where the distribution over phrases is conditioned on a given vector of terms $\mathbf{t}$ appearing in the *document* containing the phrase. We note that given sufficient contextual information, word senses may become unambiguous from the context, obviating some of the need for non-NL URIs.

5.2 A Market for Knowledge

We expect that different KBs, constructed using different methods and for different purposes, will specialize in different knowledge. As a result, a new category of services need to be developed that can advise a KB about which other KBs can answer which questions. How should a KB combine evidence across other KBs? And what incentive structures will reward KBs for providing high-quality output, and help WIE systems to focus on extracting new knowledge that is helpful for other KBs? When a KB extracts new knowledge, should it disseminate this to interested KBs rather than waiting to be queried? New mechanisms would need to be designed for these purposes.

5.3 Clarifying the NL Objective

While the experiments in Section 4 show that the perplexity measure correlates well with WIE performance for a particular class of models (HMMs), in general the NL Objective requires further refinement.

In the limit, a KB that performs sufficiently well according to the NL Objective will necessarily contain a vast, useful body of knowledge. However, in the nearer-term, optimizing the NL Objective in terms of standard metrics like perplexity has the potential to be counter-productive. As a specific example, extractors based on trigram models can actually be shown to be *less* accurate for WIE than HMMs, but the trigrams achieve better perplexity scores in modeling $P(\mathbf{w})$ through the use of careful backoff and smoothing techniques.

Thus, new metrics must be developed for evaluating distributions $P(\mathbf{w})$. We desire metrics that vary monotonically with the "knowledge content" of a KB. The ideal metric would penalize *semantic* errors, rather than (for example) rewarding particularly precise probability estimates on common phrases. The recent "adversarial evaluation" approach for NL models represents one promising direction [31].

It is also important to note that even with trillions of training examples, language model performance will typically remain an imperfect measure of KB capabilities. For example, the NL Objective may not reflect whether a KB can perform arithmetic (plenty of reasonable sums, such as "25 plus 329," never occur even in a corpus as large as the Web). Characterizing what types of knowledge are, and are not, reflected by the NL Objective is an important task.

5.4 Scaling Language Model Training

The NL Objective suggests that building more predictive language models is an important direction for WIE. In particular, we require language models that can handle potentially lengthy queries—i.e. models that infer which strings are likely to occur, even if the strings never appear on the Web. Latent variable models such as HMMs are one step toward such a model; other promising avenues include deep neural network language models [32] and recent models that include language compositionality [33, 34].

While parallel training techniques have been developed for many models [35, 36], training sophisticated models on large corpora with large vocabularies is an ongoing challenge. Can we develop new techniques that actively select *human* input to improve the models? In some settings, carefully selecting informative input can dramatically reduce the amount of training required [37], but these techniques have not been applied to modern statistical language models. A related direction involves developing new learning approaches that do not iterate over the entire corpus, but instead learn from selected *statistics* computed over the data (e.g. [38]).

6. CONCLUSION

We proposed a framework that utilizes natural language for integrating, evaluating, and optimizing extracted knowledge bases (KBs). In the NL protocol, KBs exchange knowledge by asking and answering queries about probability distributions over language. The NL Objective evaluates and optimizes KBs in terms of their ability to accurately estimate probabilities over language. Several research challenges remain. Our next steps include implementing the NL protocol over existing extracted KBs, and evaluating its effectiveness experimentally.

Acknowledgments

This work was supported in part by NSF Grants IIS-1016754 and IIS-1065397, and DARPA contract D11AP00268. We thank the anonymous reviewers for helpful comments.

References

[1] Cynthia Matuszek Michael, Michael Witbrock, Robert C. Kahlert, John Cabral, Dave Schneider, Purvesh Shah, and Doug Lenat. Searching for common sense: Populating cyc from the web. In *In Proceedings of the Twentieth National Conference on Artificial Intelligence*, pages 1430–1435, 2005.

[2] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[3] Kenneth D Forbus, Christopher Riesbeck, Lawrence Birnbaum, Kevin Livingston, Abhishek Sharma, and Leo Ureel. Integrating natural language, knowledge representation and reasoning, and analogical processing to learn by reading. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 22, page 1542. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[4] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the Web. In *Procs. of IJCAI*, 2007.

[5] Wolfgang Gatterbauer, Paul Bohunsky, Marcus Herzog, Bernhard Krüpl, and Bernhard Pollak. Towards domain-independent information extraction from web tables. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 71–80, New York, NY, USA, 2007. ACM.

[6] Michael J. Cafarella, Alon Y. Halevy, Daisy Z. Wang, Eugene W. 0002, and Yang Zhang. Webtables: exploring the power of tables on the web. *PVLDB*, 1(1):538–549, 2008.

[7] Fei Wu and Daniel S. Weld. Automatically refining the wikipedia infobox ontology. In *Proc. of WWW*, 2008.

[8] F.M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Procs. of WWW*, 2007.

[9] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[10] Sören Auer and Jens Lehmann. What have innsbruck and leipzig in common? extracting semantics from wiki content. In *Proc. of ESWC*, 2007.

[11] James Fan, David Ferrucci, David Gondek, and Aditya Kalyanpur. Prismatic: Inducing knowledge from a large scale lexicalized relation resource. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 122–127. Association for Computational Linguistics, 2010.

[12] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

[13] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Procs. of the 14th International Conference on Computational Linguistics*, pages 539–545, Nantes, France, 1992.

[14] Doug Downey, Oren Etzioni, and Stephen Soderland. Analysis of a probabilistic model of redundancy in unsupervised information extraction. *Artificial Intelligence*, 174(11):726 – 748, 2010.

[15] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. Organizing and searching the world wide web of facts - step one: The one-million fact extraction challenge. In *AAAI 2006*. AAAI Press, 2006.

[16] Fei Wu, Raphael Hoffmann, and Daniel S. Weld. Information extraction from wikipedia: moving down the long tail. In *Proc. of KDD*, 2008.

[17] Fei Wu and Daniel S. Weld. Autonomously semantifying wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50, New York, NY, USA, 2007. ACM.

[18] Hector Gonzalez, Alon Y Halevy, Christian S Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, and Jonathan Goldberg-Kidon. Google fusion tables: web-centered data management and collaboration. In *Proceedings of the 2010 international conference on Management of data*, pages 1061–1066. ACM, 2010.

[19] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, 2002.

[20] L.K. Schubert and M.H. Tong. Extracting and evaluating general world knowledge from the brown corpus. In *Proc. of the HLT/NAACL Workshop on Text Meaning*, 2003.

[21] AnHai Doan and Alon Y. Halevy. Semantic-integration research in the database community. *AI Mag.*, 26(1):83–94, 2005.

[22] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pages 1265–1266. ACM, 2008.

[23] O. Medelyan and C. Legg. Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense. In *Proc. of WIKIAI*, 2008.

[24] D. Downey, A. Ahuja, and M. Anderson. Learning to integrate relational databases with wikipedia. In *Proc. of WIKIAI*, 2009.

[25] Thomas Lin, Oren Etzioni, et al. Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 84–88. Association for Computational Linguistics, 2012.

[26] Z. Harris. Distributional structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. New York: Oxford University Press, 1985.

[27] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Methods for Exploring and Mining Tables on Wikipedia. In *Proceedings of the ACM SIGKDD Interactive Data Exploration and Analytics (IDEA).* ACM, 2013.

[28] Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Alan Ritter, Stefan Schoenmackers, et al. Machine reading at the university of washington. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 87–95. Association for Computational Linguistics, 2010.

[29] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Automated Knowledge Base Construction (AKBC): The 3rd Workshop on Knowledge Extraction at CIKM*, 2013.

[30] Fei Huang, Arun Ahuja, Doug Downey, Yi Yang, Yuhong Guo, and Alexander Yates. Learning Representations for Weakly Supervised Natural Language Processing Tasks. *Computational Linguistics*, xx:yy, 2013.

[31] Noah A Smith. Adversarial evaluation for models of natural language. *arXiv preprint arXiv:1207.0245*, 2012.

[32] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[33] Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429, 2010.

[34] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

[35] Jason Wolfe, Aria Haghighi, and Dan Klein. Fully distributed em for very large datasets. In *ICML*, 2008.

[36] Yi Yang, Alexander Yates, and Doug Downey. Overcoming the memory bottleneck in distributed training of latent variable models of text. In *Proceedings of NAACL-HLT*, pages 579–584, 2013.

[37] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[38] Michael Lucas and Doug Downey. Scaling semi-supervised naive bayes with feature marginals. In *Proceedings of ACL*, 2013.