Adding Distributional Semantics to Knowledge Base Entities through Web-scale Entity Linking

Matthew Gardner
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
mg1@cs.cmu.edu

Abstract

Web-scale knowledge bases typically consist entirely of predicates over entities. However, the distributional properties of how those entities appear in text are equally important aspects of knowledge. If noun phrases mapped unambiguously to knowledge base entities, adding this knowledge would simply require counting. The many-to-many relationship between noun phrase mentions and knowledge base entities makes adding distributional knowledge about entities difficult. In this paper, we argue that this information should be explicitly included in web-scale knowledge bases. We propose a generative model that learns these distributional semantics by performing entity linking on the web, and we give some preliminary results that point to its usefulness.

1 Introduction

Recent work in automatically creating web-scale knowledge bases (like YAGO, Freebase, and NELL) has focused on extracting properties of concepts and entities that can be expressed as n-ary relations (Suchanek et al., 2007; Bollacker et al., 2008; Carlson et al., 2010b). Examples might be Athlete(`Michael Jordan 1`), Professor(`Michael Jordan 2`), PlaysForTeam(`Michael Jordan 1`, `Chicago Bulls`), and UniversityFaculty(`UC Berkeley`, `Michael Jordan 2`). The task of the knowledge extraction algorithm is to find new instances of these relations given some training examples, perhaps while jointly determining the set of relevant entities.

While these knowledge extraction approaches have focused on relational knowledge, knowing how `UC Berkeley` appears distributionally in text is also an important aspect of the entity that is potentially useful in a variety of tasks. For example, Peñas and Hovy (2010) showed that a collection of distributional knowledge about football entities helped in interpreting noun compounds like "Young touchdown pass." Haghighi and Klein (2010) used distributional information about entity types to achieve state-of-the-art coreference resolution results. It has long been known that word sense disambiguation and other tasks are best solved with distributional information (Firth, 1957), yet this information is lacking in web-scale knowledge bases.

The primary reason that distributional information has not been included in web-scale knowledge bases is the inherent ambiguity of noun phrases. Knowledge bases typically aim to collect facts about entities, not about noun phrases, but distributional information is only easily obtained for noun phrases. In order to add distributional semantics to knowledge base entities, we must perform *entity linking*, determining which entity any particular noun phrase in a document refers to, at web scale.

We suggest that distributional semantics should be included explicitly in web-scale knowledge bases, and we propose a generative model of entity linking that learns these semantics from the web. This would both enrich the representation of entities in these knowledge bases and produce better data for further relational learning. In the next section, we frame this idea in the context of prior work. In Section 3, we describe a model that learns distributional

Proc. of the Joint Workshop on Automatic Knowledge Base Construction & Web-scale Knowledge Extraction (AKBC-WEKEX), pages 46–51,
NAACL-HLT, Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

semantics for the set of entities in a knowledge base in the context of an entity linking task. Finally, in Section 4 we conclude.

2 Related Work

Our work builds off of a few related ideas. First, Haghighi and Klein (2010) presented a coreference resolution system that had at its core a set of distributional semantics over entity types very similar to what we propose. For each of a set of entity types (like Person, Organization, and Location) and each of a set of properties (like "proper head," "common head," "subject of verb"), they have a distribution over values for that property. People are thus more likely to have "Steve" or "John" as noun modifiers, while Organizations are more likely to have "Corp." as proper heads.

Their system learned these distributions in a semi-supervised fashion, given a few seed examples to their otherwise unsupervised coreference model. Their system did not, however, have any notion of global *entities*; they had global *types* whose parameters were shared across document-specific entities. Every time they saw the noun phrase "Barack Obama" in a new document, for example, they created a new entity of type "Person" for the mentions in the document. Even though they did not model individual entities, their system achieved state-of-the-art coreference resolution results. We believe that their modeling of distributional semantics was key to the performance of their model, and we draw from those ideas in this paper.

Our proposal is also very similar to ideas presented by Hovy (2011). Hovy describes a "new kind of lexicon" containing both relational information traditionally contained in knowledge bases and distributional information very similar to that used in Haghighi and Klein's coreference model. Each item in this "new lexicon" is represented as a set of distributions over feature values. The lexical entry for "dog," for example, might contain a feature "name," with "Spot" and "Lassie" receiving high weight, and a feature "agent-of," with highly probable values "eat," "run," and "bark." While Hovy has presented this vision of a new lexicon, he has left as open questions how to actually construct it, and how compositionality, dependence, and logical operators can function efficiently in such a complex system.

Peñas and Hovy (2010) have shown how a very small instance of a similar kind of lexicon can perform well at interpreting noun compounds, but they needed to resort to a severely restricted domain in order to overcome the challenges of constructing the lexicon. Because they only looked at a small set of news articles about football, they could accurately assume that all mentions of the word "Young" referred to a single entity, the former San Francisco 49ers quarterback. At web scale, such assumptions quickly break down.

There has been much recent work in distantly supervised relation extraction, using facts from a knowledge base to determine which sentences in a corpus express certain relations in order to build relation classifiers (Hoffmann et al., 2011; Riedel et al., 2010; Mintz et al., 2009). This work depends on first performing entity linking, finding sentences which contain pairs of knowledge base entities. Typically, this linking has been a simple string-matching heuristic, a noisy alignment that throws away a lot of useful information. Using coreference resolution after a noisy alignment can help to mitigate this issue (Gabbard et al., 2011), but it is still mostly a heuristic matching. A benefit of our approach to adding distributional semantics to web-scale knowledge bases is that in the process we will create a large entity-disambiguated corpus that can be used for further relational learning.

3 Entity Linking

We add distributional semantics to knowledge base entities through performing entity linking. Specifically, given a knowledge base and a collection of dependency parsed documents, entity linking maps each noun phrase in the document collection to an entity in the knowledge base, or labels it as unknown (a deficiency we will address in future work). Our model does this by learning distributions over dependency link types and values for each entity in the knowledge base. These distributions are both the features that we use for entity linking and the distributional semantics we aim to include in the knowledge base.

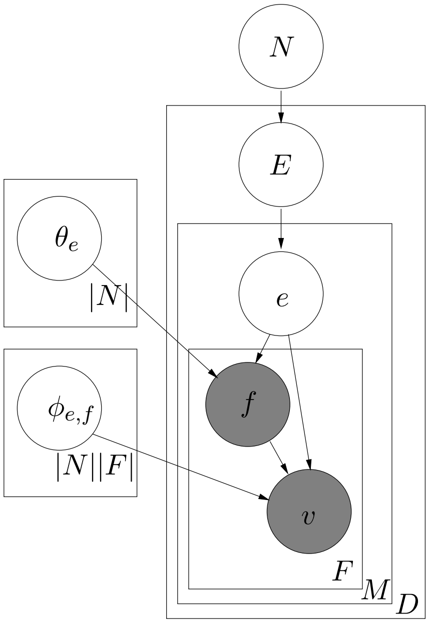

Figure 1: Graphical model for entity linking.

3.1 Model Structure

The model we propose is similar in structure to hierarchical models like latent Dirichlet allocation (Blei et al., 2003) or hierarchical Dirichlet processes (Teh et al., 2006). Instead of the "topics" of those models, we have entities (i.e., one "topic" in the model for every entity in the knowledge base, plus one "unknown" topic), and instead of modeling individual words, we model entity mentions in the document.

The generative story of the model is as follows. First, given a set of entities from a knowledge base, fix a Dirichlet prior N over them, and draw a set of multinomial parameters $\phi_{e,f}$ and θ_e for each entity from a set of Dirichlet priors α and β. Next, for each of D documents, draw a multinomial distribution over entities E appearing in that document from N. Then for each of M mentions in the document, draw from E an entity e to which that mention refers. Given the entity e, draw a set of F feature types f from θ_e. For each feature type f, draw a feature value v from the distribution $\phi_{e,f}$ corresponding to the entity e and the feature type f. This model is shown graphically in Figure 1.

We chose a generative model with multinomial distributions instead of other options because we want the resultant distributions ϕ and θ to be immediately interpretable and usable in other models, as the intent is that they will be stored as part of the knowledge in the knowledge base. Also, we intend to extend this model to allow for the creation of new entities, a relatively easy extension with a model of this form.

3.2 Features

Here we describe in more detail what we use as the features f in the model. These features and their corresponding parameters ϕ and θ constitute the distributional information that we propose to include in web-scale knowledge bases, and they aim to capture the way knowledge-base entities tend to appear in text.

The features we propose are the set of Stanford dependency labels that attach to the head word of each mention, with the values being its dependents or governors. We also have features for the head word of the mention, whether it is a proper noun, a common noun, or a pronoun. We keep track of the direction of the dependencies by prepending "gov-" to the dependency label if the mention's head word is governed by another word, and we stem verbs. For example, in the sentence "Barack Obama, president of the United States, spoke today in the Rose Garden," the mention "Barack Obama" would have the following features:

Feature	Value
proper-head	Obama
nn	Barack
gov-nsubj	speak
appos	president

When there are deterministically coreferent mentions, as with appositives, we combine the features from both mentions in preprocessing.

We note here also that we use dependency links as features over which to learn distributional semantics because they are the deepest semantic representation that current tools will allow us to use at web scale. We would like to eventually move from dependency links to semantic roles, and to include relations expressed by the sentence or paragraph as features in our model. One possible way of doing that is to use something like ReVerb (Fader et al., 2011), setting its output as the value and an unobserved relation in the knowledge base as the feature type. This would learn distributional information about the textual expression of relations directly, which would also be very useful to have in web-scale knowledge bases.

3.3 Inference

Inference in our model is done approximately in a MapReduce sampling framework. The map tasks sample the entity variables for each mention in a document, sequentially. The entity variables are constrained to either refer to an entity already seen in the document, or to a new entity from the knowledge base (or unknown). Sampling over the entire knowledge base at every step would be intractable, and so when proposing a new entity from the knowledge base we only consider entities that the knowledge base considers possible for the given noun phrase (e.g., NELL has a "CanReferTo" relation mapping noun phrases to concepts (Krishnamurthy and Mitchell, 2011), and Freebase has a similar "alias" relation). Thus the first mention of an entity in a document must be a known alias of the entity, but subsequent mentions can be arbitrary noun phrases (e.g., "the college professor" could not refer to `Michael Jordan 2` until he had been introduced with a noun phrase that the knowledge base knows to be an alias, such as "Michael I. Jordan"). This follows standard journalistic practice and aids the model in constraining the "topics" to refer to actual knowledge base entities.

The reduce tasks reestimate the parameters for each entity by computing a maximum likelihood estimate given the sampled entity mentions from the map tasks. Currently, there is no parameter sharing across entities, though we intend to utilize the structure of the knowledge base to tie parameters across instances of the same category in something akin to a series of nested Dirichlet processes.

While we have not yet run experiments with the model at web scale, it is simple enough that we are confident in its scalability. Singh et al. and Ahmed et al. have shown that similarly structured models can be made to scale to web-sized corpora (Singh et al., 2011; Ahmed et al., 2012).

3.4 Evaluation

Evaluating this model is challenging. We are aiming to link *every noun phrase* in every document to an entity in the knowledge base, a task for which no good dataset exists. It is possible to use Wikipedia articles as labeled mentions (as did Singh et al. (2011)), or the word sense labels in the OntoNotes corpus (Weischedel et al., 2011), though these require a mapping between the knowledge base and Wikipedia entities or OntoNotes senses, respectively. The model also produces a coreference decision which can be evaluated. These evaluation methods are incomplete and indirect, but they are likely the best that can be hoped for without a labor-intensive hand-labeling of large amounts of data.

3.5 Preliminary Results

We do not yet have results from evaluating this model on an entity linking task. However, we do have preliminary distributional information learned from 20,000 New York Times articles about baseball. Some of the distributions learned for the New York Mets baseball team are as follows.

gov-nsubj	gov-poss
had: 0.040	manager: .088
have: 0.035	president: .032
won: 0.028	clubhouse: .024
lost: 0.026	victory: .024
got: 0.018	baseman: .020
scored: 0.015	coach: .019

These distributions themselves are inherently useful for classification tasks—knowing that an entity possesses managers, presidents, basemen and coaches tells us a lot about what kind of entity it is. The learning system for the NELL knowledge base currently uses distributions over noun phrase contexts (a few words on either side) to learn information about its concepts (Carlson et al., 2010a). The results of this model could provide much better data to NELL and other learning systems, giving both more structure (distributions over dependency links instead of windowed contexts) and more refined information (distributions over concepts directly, instead of over noun phrases) than current data sources.

4 Conclusion

We have argued for the inclusion of distributional semantics directly in web-scale knowledge bases. This is more difficult than simple counting because of the inherent ambiguity in the noun phrase to entity mapping. We have presented a model for obtaining this distributional knowledge for knowledge base entities (instead of for ambiguous noun phrases) by performing entity linking at web scale. While producing useful distributional knowledge about entities, this work will also provide much richer data sources to traditional relation extraction algorithms. Though our work is still preliminary and there are challenges to be overcome, the primary purpose of this paper is to argue that this research direction is feasible and worth pursuing. A knowledge base that includes both properties about entities and distributional knowledge of how those entities appear in text is much more useful than a knowledge base containing facts alone.

Acknowledgments

We thank Jayant Krishnamurthy and Tom Mitchell for helpful conversations, the anonymous reviewers for their comments, and the Intel Science and Technology Center on Embedded Computing for financial support.

References

A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A.J. Smola. 2012. Scalable inference in latent variable models. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 123–132. ACM.

David M. Blei, Andrew Y. Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. 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*, pages 1247–1250. ACM.

A. Carlson, J. Betteridge, R.C. Wang, E.R. Hruschka Jr, and T.M. Mitchell. 2010a. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 101–110. ACM.

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

A. Fader, S. Soderland, and O. Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

J.R. Firth, 1957. *A synopsis of linguistic theory 1930–1955*, pages 1–32. Philological Society, Oxford.

R. Gabbard, M. Freedman, and R. Weischedel. 2011. Coreference for learning to extract relations: yes, virginia, coreference matters. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 288–293. Association for Computational Linguistics.

A. Haghighi and D. Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 385–393. Association for Computational Linguistics.

R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D.S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

E.H. Hovy. 2011. Toward a new semantics: Merging propositional and distributional information. Presentation at Carnegie Mellon University.

J. Krishnamurthy and T.M. Mitchell. 2011. Which noun phrases denote which concepts? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 570–580.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1011. Association for Computational Linguistics.

A. Peñas and E. Hovy. 2010. Filling knowledge gaps in text for machine reading. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 979–987. Association for Computational Linguistics.

S. Riedel, L. Yao, and A. McCallum. 2010. Modeling relations and their mentions without labeled text. *Machine Learning and Knowledge Discovery in Databases*, pages 148–163.

S. Singh, A. Subramanya, F. Pereira, and A. McCallum. 2011. Large-scale crossdocument coreference using distributed inference and hierarchical models. *Association for Computational Linguistics: Human Language Technologies (ACL HLT)*.

F.M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, et al. 2011. Ontonotes release 4.0. Linguistic Data Consortium.