Topic Models for Corpus-centric Knowledge Generalization

Benjamin Van Durme Daniel Gildea

The University of Rochester
Computer Science Department
Rochester, NY 14627

Technical Report 946

June, 2009

Abstract

Many of the previous efforts in generalizing over knowledge extracted from text have relied on the use of manually created word sense hierarchies, such as WordNet. We present initial results on generalizing over textually derived knowledge, through the use of the LDA topic model framework, as the first step towards automatically building corpus specific ontologies.

This work was funded by a 2008 Provost's Multidisciplinary Award from the University of Rochester, and NSF grants IIS-0328849 and IIS-0535105.

1 Introduction

Many of the previous efforts in generalizing knowledge extracted from text (e.g., Suchanek et al. (2007), Banko and Etzioni (2007), Paşca (2008), and Van Durme et al. (2009)) have relied on the use of manually created word sense hierarchies, such as WordNet. Unfortunately, as these hierarchies are constructed based on the intuitions of lexicographers or knowledge engineers, rather than with respect to the underlying distributional frequencies in a given corpus, the resultant knowledge collections may be less applicable to language processing tasks.

While previous work (such as by Cimiano et al. (2005), or Liakata and Pulman (2008)) has explored the use of *automatically* constructed taxonomies, these methods tend to make strong assumptions about word sense, where each surface token is forced to have just a single underlying meaning.

Our eventual goal is the construction of corpus-specific, probabilistic ontologies, that allow for multiple word senses, and whose structure reflects the distributional semantics found in text. Such an ontology will allow for robust generalization of extracted knowledge, providing a *softer* alternative to the brittleness of traditional knowledge bases that has thus far largely prevented their use in real world textual applications.

We present here the first step towards this goal: the application of the LDA Topic Model framework of Blei et al. (2003) in order to derive a set of classes based on underlying semantic data: a collection of automatically extracted propositions conveying general world knowledge.

2 Generalizing Knowledge

2.1 Background Knowledge

Our experiments are based on the data of Van Durme et al. (2009), consisting of a large collection of general knowledge extracted from the British National Corpus (BNC) through use of those authors' KNEXT system. KNEXT operates through a set of manually constructed semantic rules, applied to syntactic parse trees in the style of the Penn Treebank.

Our problem specification starts from that of Van Durme et al.: given a collection of individual propositions, e.g., A MALE MAY BUILD A HOUSE, automatically extracted from some corpus, construct conditional statements that can be viewed as stronger claims about the world, e.g., IF A MALE BUILDS SOMETHING, THEN IT IS PROBABLY A STRUCTURE, A BUSINESS, OR A GROUP. We revise this problem to be: given such a collection, construct conditional *probability distributions* that reflect the likelihood of a proposition being expressed with a particular argument, in text similar to that of the original corpus. For example, if we were to sample values for X, conditioned on a *propositional template* such as, A MALE MAY BUILD X, we might see examples such as: A HOUSE, SOME STEPS, AN ANIMAL, A HOUSE, A LYRICISM, A CHURCH, KARDAMILI, SERVICES, A CAGE, A CAMP, ...[1]

[1] These are the actual first 10 arguments sampled from a 100 topic model during development.

2.2 Model Description

Let each propositional template or *contextual relation* be indexed by $r \in \mathcal{R} = \{1, ..., M\}$, limited in this work to having a single free argument $a \in \mathcal{A}$, where $\mathcal{A}$ is finite.

Assume some set of observations, taking the form of pairs ranging over $\mathcal{R} \times \mathcal{A}$. Then the list of non-unique arguments seen occurring with relation r is written as, $\boldsymbol{a}_r = (a_{r1}, ..., a_{rN_r}) \in \mathcal{A}^{N_r}$. For example, the indices in a pair $\langle r, a \rangle$ might correspond to: $\langle$A MALE MAY BUILD, A HOUSE$\rangle$, while the indices in an argument list, $\boldsymbol{a}_r$, might correspond to:(A HOUSE, A HOUSE, A HOUSE, SOME STEPS, ...).

We are concerned here with $\Pr(a \mid r)$: the probability of an argument, a, given a relation, r. Let

$$c_r(a) = \sum_{i=1}^{N_r} \delta(a = a_{ri})$$

be the number of times argument a, was observed with r. The maximum likelihood estimate (MLE) is then

$$\hat{\Pr}(a \mid r) = \frac{c_r(a)}{N_r}.$$

Assume the observation set is sparse, where the resultant MLE may incorrectly assign zero mass to events that have low (but non-zero) probability. Further assume that the distributions associated with distinct relations are not independent. For example, I expect the context A FIREFIGHTER MAY EAT X to have an argument pattern similar to A SECRET SERVICE AGENT MAY EAT X. To capture this intuition I introduce a set of hidden *topics*, which here represent semantic classes as probability distributions over unique arguments. Under this model, we imagine a given argument is generated by first selecting some topic based on the relation, and then selecting an argument based on the topic.

Where $z \in \mathcal{Z} = \{1, ..., T\}$ is a set of topics, let $\phi_z(a) = \Pr(a \mid z)$ be the probability of an argument given a topic, and $\theta_r(z) = \Pr(z \mid r)$ be the probability of a topic given a context. Both θ_r and ϕ_z represent multinomial distributions, whose parameters we will estimate based on training data.

The revised formula for the probability of a given r becomes

$$\Pr(a \mid r) \quad = \quad \sum_{z \in \mathcal{Z}} \phi_z(a)\theta_r(z) = \sum_{z \in \mathcal{Z}} \Pr(a \mid z) \Pr(z \mid r).$$

I use here the Latent Dirichlet Allocation (LDA) framework[2] introduced by Blei et al. (2003), a generative model for describing the distribution of elements within, e.g., a document collection. Using the terminology of this model, documents are represented as a weighted mixture of underlying topics, with each topic representing a multinomial distribution over the vocabulary (see Figure 1).

[2]Specifically, I use the *smoothed* LDA model of Blei et al. (2003), where the topic multinomials, ϕ, are also taken to be draws from a Dirichlet prior, in addition to the document multinomials, θ.

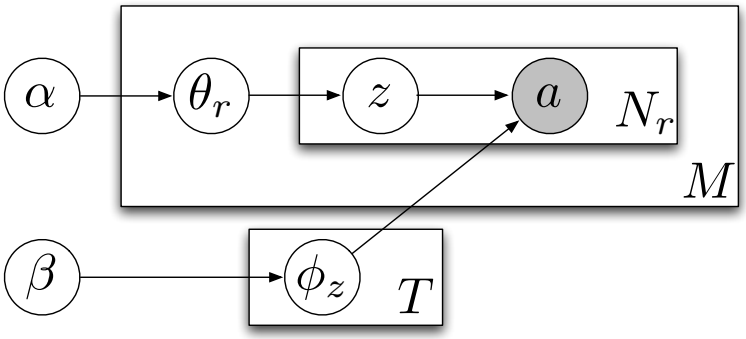

Figure 1: The smoothed LDA Model of Blei et al. (2003), in *plate notation*: (non)shaded circles represent (un)observed variables, arrows represent dependence, and boxes with some term x in the lower right corner represents a process repeated x times.

This model assumes observations are generated by a process in which each document's topic distribution θ_r is first sampled from an underlying symmetric[3] Dirichlet distribution with parameter α and then each word of the document is generated conditioned on θ_r and the multinomial distributions represented by each topic. Those topic distributions are themselves taken as draws from a symmetric Dirichlet distribution with parameter β:

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\alpha), \\ z \mid \theta_r &\sim \text{Multinomial}(\theta_r), \\ \phi &\sim \text{Dirichlet}(\beta), \\ a \mid z, \phi &\sim \text{Multinomial}(\phi_z). \end{aligned}$$

With respect to topic model terminology, $\mathcal{R}$ may be considered a set of indices over *documents*, each associated with a list of observed *words*. In this case N_r is the *size* of document r.

2.3 Parameter Inference

Based on observations, we must estimate $T + M$ distinct multinomial distributions represented by the probability functions $\phi_1, ..., \phi_T$ and $\theta_1, ..., \theta_M$.

Parameter inference was carried out using the Gibbs sampling procedure described by Steyvers and Griffiths (2007), the implementation of which is given in Figure 2.

[3] A Dirichlet distribution over an n-dimensional multinomial is defined by n separate parameters, α_1 to α_n. A *symmetric* Dirichlet distribution constrains all n parameters to equal a single value, α.

Given:

$\mathcal{R}$: set of relations

$\boldsymbol{a}_1, ..., \boldsymbol{a}_M$: arguments seen for each relation

$\mathcal{C}^{\mathcal{A}\mathcal{Z}}$: an $|\mathcal{A}| \times T$ matrix

$\mathcal{C}^{\mathcal{R}\mathcal{Z}}$: an $M \times T$ matrix

$\mathcal{C}^{\mathcal{Z}}$: a vector mapping topic ids to argument counts

$\mathcal{C}^{\mathcal{R}}$: a vector mapping relation ids to argument counts, i.e., $|\boldsymbol{a}_r|$

$c(r, a, z)$: a map from a relation, argument, topic triple to a count

Parameters:

T : number of topics

α, β : Dirichlet parameters

n : number of iterations

Initialize:

Initialize $c(\cdot, \cdot, \cdot)$, and the cells of $\mathcal{C}^{\mathcal{Z}}$, $\mathcal{C}^{\mathcal{A}\mathcal{Z}}$, and $\mathcal{C}^{\mathcal{R}\mathcal{Z}}$ to 0

For each $r \in \mathcal{R}$:

For each $a \in \boldsymbol{a}_r$:

Draw a topic id, z, from (1, ..., T), uniformly at random

Increment $c(r, a, z)$, $\mathcal{C}^{\mathcal{Z}}_{z}$, $\mathcal{C}^{\mathcal{A}\mathcal{Z}}_{az}$, and $\mathcal{C}^{\mathcal{R}\mathcal{Z}}_{rz}$

Inline-Function SAMPLE:

Let s be 0

Let $\mathcal{V}$ be a vector of size T

For t from 1 to T:

Let $\mathcal{V}_t$ be $\left(\frac{\mathcal{C}^{\mathcal{A}\mathcal{Z}}_{az} + \beta}{\mathcal{C}^{\mathcal{Z}}_{z} + |\mathcal{A}|\beta}\right)\left(\frac{\mathcal{C}^{\mathcal{R}\mathcal{Z}}_{rz} + \alpha}{\mathcal{C}^{\mathcal{R}}_{r} + T\alpha}\right)$

Increment s by $\mathcal{V}_t$

Normalize $\mathcal{V}$ by s

Draw a topic id, z', from (1, ..., T), at random according to the multinomial $\mathcal{V}$

Algorithm:

For e from 1 to n:

For each $r \in \mathcal{R}$:

For each unique $a \in \boldsymbol{a}_r$:

For each z such that $c(r, a, z) > 0$:

Let x be $c(r, a, z)$

For i from 1 to x:

Decrement $c(r, a, z)$, $\mathcal{C}^{\mathcal{Z}}_{z}$, $\mathcal{C}^{\mathcal{R}}{}_{r}$, $\mathcal{C}^{\mathcal{A}\mathcal{Z}}_{az}$, and $\mathcal{C}^{\mathcal{R}\mathcal{Z}}_{rz}$

SAMPLE new topic id, z'

Increment $c(r, a, z')$, $\mathcal{C}^{\mathcal{Z}}_{z'}$, $\mathcal{C}^{\mathcal{R}}_{r}$, $\mathcal{C}^{\mathcal{A}\mathcal{Z}}_{az'}$, and $\mathcal{C}^{\mathcal{R}\mathcal{Z}}_{rz'}$

Figure 2: Gibbs sampling procedure used for parameter inference, where snapshots of count matrices $\mathcal{C}^{\mathcal{A}\mathcal{Z}}$ and $\mathcal{C}^{\mathcal{R}\mathcal{Z}}$ may be taken at even intervals following burn-in, and then averaged and normalized to provide estimates of $\phi_1, ..., \phi_T$ and $\theta_1, ..., \theta_M$.

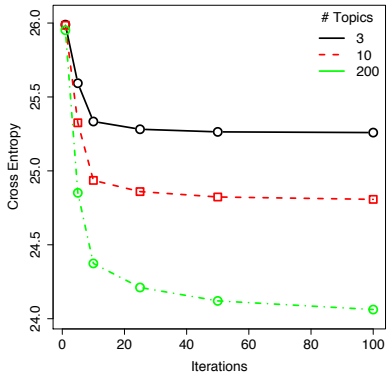

Figure 3: Cross entropy of topic models with 3, 10 and 200 topics, evaluated on held out data, presented as a function of iterations of Gibbs sampling.

3 Experiments

3.1 Data

Our dataset was constructed from the propositional templates described by Van Durme et al. (2009). We considered just those templates with a frequency of 10 or more, which gave approximately 140,000 unique templates (documents), over a vocabulary of roughly 500,000 unique arguments (words). We held out 5% of this collection, randomly sampled, to use for evaluation, with the rest being used to train models of varying number of underlying topics.

3.2 Building Models

Following suggestions by Steyvers and Griffiths (2007), we fixed $\alpha = 50/T$ and $\beta = 0.01$, then constructed models across a variety of fixed values for T. As topic distributions showed little variation after burn-in, for the exploratory work reported here we simply took θ and ϕ as the final configurations of the model at the final iteration of the chain.

Figure 3 shows the cross entropy of three models on our held out data, as a function of iteration. As seen, more underlying topics allow for better predictions of argument patterns in unseen text. Increasing the number of topics significantly beyond 200 was computationally problematic for this dataset; further investigation along these lines is a point for future work.

For the resultant models, examples of the most probable arguments given a topic can be seen in Table 1. Table 2 gives examples for the most probable templates to be observed in the training data, once the topic is fixed.

3.3 Evaluating Models

From the training corpus, 100 templates were sampled based on their frequency. Each template was evaluated according to the 5 point scale used by Van Durme et al. (2009): a score of 1 corresponds to

0		1		6	
⟨a.d end.n⟩	AN END	(k ⟨plur person.n⟩)	PEOPLE	⟨sm.q ⟨plur eye.n⟩⟩	EYES
⟨a.d part.n⟩	A PART	(k ⟨plur child.n⟩)	CHILDREN	⟨a.d head.n⟩	A HEAD
⟨a.d problem.n⟩	A PROBLEM	(k ⟨plur woman.n⟩)	WOMEN	⟨a.d life.n⟩	A LIFE

Table 1: From the model with 10 underlying topics, the 3 most probable arguments from topics 0, 1 and 6. Presented both in the underlying *pseudo logical form* representation used by KNEXT, along with the associated English verbalization.

27	62	108
AN INTEREST CAN BE IN X	X MAY HAVE A REGION	A PERSON MAY HEAR X
X MAY UNDERGO A TEACHING	X MAY HAVE AN AREA	X MAY RING
A BOOK CAN BE ON X	X MAY HAVE A COAST	X MAY HAVE A SOUND

Table 2: From the model with 200 underlying topics, the 3 most probable templates when conditioned on topics 27, 62 and 108.

a template that can be combined with some argument in order to form a *"Reasonably clear, entirely plausible general claim"*, while a score of 5 corresponds to bad propositions. From the sample, templates such as X MAY ASK were judged poorly, as they appear to be missing a central argument (e.g., ... A QUESTION, or ... A PERSON), while a few were given low scores because of noise in KNEXT's processing or the underlying corpus (e.g., X MAY HM). Examples of high quality templates can be seen as part of Table 3.

Average assessment of the first 100 templates was 1.99, suggesting that for the majority of our data there does exist at least one argument for which the fully instantiated proposition would be a reasonable claim about the world. The distribution of assessments may be seen in Figure 4.

A further 38 propositions were sampled until we had 100 templates judged as 2 or better (this led to a final mean of 1.94 over the extended sample of 138 templates).

For each of these high quality templates, one argument was drawn from each of three models. These arguments where then used to instantiate complete propositions, presented for evaluation absent the information of which argument came from which model. Table 3 gives three such (template, argument) pairs, along with the assessed quality. In Table 4 we see that just as cross entropy decreased with the addition of more topics, the quality of sampled arguments improves as well.

	3		10		200	
X MAY BE COOKED	IMPLICATIONS	5	A WEB	5	FISH	1
A PERSON MAY PAY X	A DISPLAY	5	A WORKER	1	A FARE	1
A DIVERSITY CAN BE IN X	OPPORTUNITIES	1	A CENTURY	3	VOLUME	1

Table 3: Examples of propositional templates and arguments drawn randomly for evaluation, along with their judgements from 1 to 5 (lower is better).

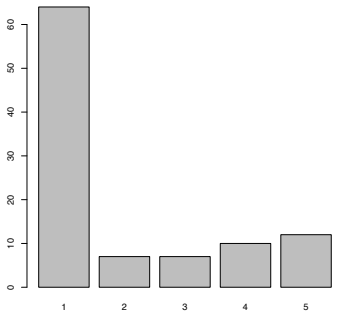

Figure 4: From a sample of 100 templates, number of those assessed at: 1 (64 of 100), 2 (7), 3 (7), 4 (10) and 5 (12).

# Topics	3	10	200
Avg. Assessment	2.39	2.09	1.73

Table 4: Starting with 100 templates assessed at 2 or better, results of drawing 100 arguments (1 per template) from models built with 3, 10 and 200 topics.

3.4 Topic Pruning per Relation

Van Durme et al. (2009) gave an algorithm that assumed the missing argument for a given propositional template was type-restricted by at most a small number of distinct categories. For example, usually if there is some X such that A PERSON MAY TRY TO FIND X, then according to evidence gathered from the BNC, we expect X to usually be either a A PERSON or A LOCATION. If we take topics to represent underlying semantic categories, then in order to enforce a similar restriction here, we need some way to constrain our model such that for each relation we keep track of at most k relevant topics. This could be achieved simply by post-processing each $\hat{\theta}_r$ by setting to 0 all non-top k topics, and re-normalizing, but this would be overly severe: the assumption is that an argument is *usually* one of a few categories, while the proposed post-processing technique would equate to a claim that it is *always* of this restricted set.

Here we investigated the effect of modifying the posterior probability such that $\Pr(a \mid r)$ is determined by using θ_r for just the relative top k, with leftover mass uniformly distributed amongst the remaining categories, weighted by the respective category probability.

Let $\mathcal{Z}_r^k$ be those $z \in \mathcal{Z}$ such that there are no more than $k-1$ such $z' \in \mathcal{Z}$ where $\theta_r(z') > \theta_r(z)$. Then we define the k-constrained probability of a given r as:

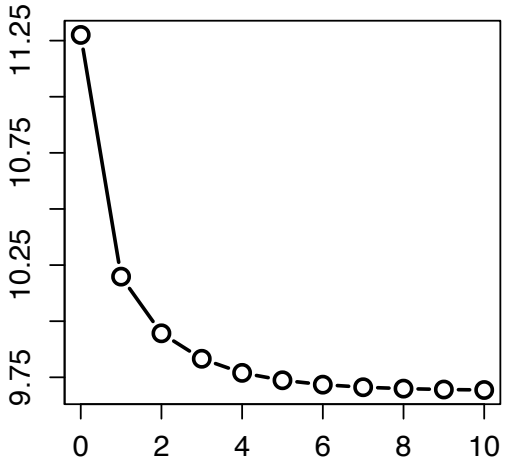

Figure 5: On a 200 topic model, for $k = 0, 1, ..., 10$, cross entropy scores on held out data using k-constrained pruning.

$$\begin{aligned}
\Pr_k(a \mid r) &= \sum_{z \in \mathcal{Z}_r^k} \phi_z(a)\theta_r(z) + \lambda_1 \left(\sum_{z' \notin \mathcal{Z}_r^k} \phi_{z'}(a) \frac{\Pr(z')}{\lambda_2} \right) \\
&= \sum_{z \in \mathcal{Z}_r^k} \phi_z(a)\theta_r(z) + \\
&\quad \left(1 - \sum_{z \in \mathcal{Z}_r^k} \theta_r(z)\right) \left(\sum_{z' \notin \mathcal{Z}_r^k} \phi_{z'}(a) \frac{\Pr(z')}{\sum_{z' \notin \mathcal{Z}_r^k} \Pr(z')} \right).
\end{aligned}$$

For a model built with 200 underlying topics, Figure 5 contains cross entropy results on held-out test data, when considering various levels of k. As seen, the majority of the predictive power for which arguments to expect for a given relation comes from the first two or three topics. If these topics are taken as (rough) semantic categories, then this agrees with the soft restriction employed by Van Durme et al. (2009) (where the parameter m in that algorithm is similar in function here to k).

As an aside, note the relationship between soft restriction, the k-constrained conditional probabilities here, and the use of an *abnormality* predicate when performing circumscription (McCarthy, 1980, 1986). McCarthy (1986) wrote: *Nonmonotonic reasoning has several uses. [... Such as] a very streamlined expression of probabilistic information when numerical probabilities, especially conditional probabilities, are unobtainable*. In Van Durme et al. (2009) the concern was that textual frequencies may not directly correlate with human beliefs about the world, in other words, that the conditional probabilities of interest (human beliefs regarding likelihood) are potentially unobtainable from the given resource (everyday text). Soft restriction, k-constrained probabilities, and the

abnormality predicate: all are aimed at getting the "main gist" correct, succinctly, while isolating the remaining as "other", or *abnormal*.

4 Related Work

Suchanek et al. (2007) performed extraction over Wikipedia, using a closed set of manually specified relations. Results were then automatically affixed to synsets within WordNet through use of the first sense heuristic.

Banko and Etzioni (2007), along with Paşca (2008), both relied on the first sense heuristic to generalize discovered knowledge by finding the synset that best covered their observations. The experiments of Banko and Etzioni were based on their TEXTRUNNER framework, as applied to a closed domain set of documents dealing with *nutrition*, while Paşca generalized over *class attributes* discovered through the use of search engine query logs. Van Durme et al. (2009) argued against reliance on the first sense heuristic when generalizing, but retained the use of WordNet.

Cimiano et al. (2005) automatically constructed ontologies based on extracted knowledge using the principles of Formal Concept Analysis (FCA), while Liakata and Pulman (2008) performed hierarchical clustering to derive a set of semantic classes as leaf nodes. Both groups focused on domain specific texts in order to minimize the problems of WSD.

Although evaluated on syntactic constructions rather than logical forms, the work of Pantel and Lin (2002) on the Clustering By Committee (CBC) algorithm is perhaps the most similar in motivation to what we've presented here. A comparative strength of our model is that it is fully generative, which should allow for principled integration into existing text processing frameworks (in particular, as a semantic component in language modeling).

Havasi (2009) explored the use of Singular Value Decomposition (SVD) techniques for clustering terms based on the semantic contexts from the Open Mind Common Sense (OMCS) project (Singh, 2002). The OMCS initiative has similar goals as the KNEXT project (as expressed by Schubert (2002)), but focuses on the use of human volunteers in the knowledge acquisition process.

Recently Brody and Lapata (2009) independently developed a semantic topic model framework for the task of word sense disambiguation. The authors built distinct models for each word, with individual topics standing for an underlying word sense.

Koo et al. (2008) presented results showing that a word hierarchy built in an unsupervised manner could be used to improve accuracy in syntactic parsing. The authors applied an agglomerative clustering technique that at each merge operation minimized overall divergence between old and new models, based on token bigram frequencies. We take this work as an example of an automatically acquired "knowledge" base being useful for NLP.

5 Conclusion

Our goal is the construction of probabilistic ontologies, with structure derived from the corpus from which the underlying knowledge was extracted. Here we have presented the first step towards that goal: the application of the LDA topic model to a large collection of automatically acquired, general

world knowledge. We have validated the applicability of this method when propositional templates are considered as *documents*, and semantic arguments being used as *words*.

By treating the knowledge generalization problem as one of constructing conditional probability distributions, we have opened the door for automatic evaluation. If one may assume the average quality of an underlying set of extracted knowledge is high, then measuring the cross entropy of constructed models against held out data serves as a powerful tool as compared to human evaluation.

Future work may investigate the use of multiple layers of such models as a semantic taxonomy, along with the use of WordNet as a source of knowledge (instead of a hard constraint). In addition, while the knowledge over which we built our models has inherent logical structure, here those propositions were used only as static semantic contexts with a single argument removed. We are actively considering the use of hierarchical models in order to capture multiple arguments, which should allow for a more natural model of the underlying semantics, and offer practical assistance in handling data sparsity. Finally, we plan to use nonparametric priors to infer the number of clusters, once the aforementioned tractability issues in building very large models is addressed.

References

Banko, M. and Etzioni, O. (2007). Strategies for Lifelong Knowledge Extraction from the Web. In *Proceedings of K-CAP*.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

Brody, S. and Lapata, M. (2009). Bayesian Word Sense Induction. In *Proceedings of EACL*.

Cimiano, P., Hotho, A., and Stabb, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Inteligence Research*.

Havasi, C. (2009). *Discovering Semantic Relations Using Singular Value Decomposition Based Techniques*. PhD thesis, Brandeis University.

Koo, T., Carreras, X., and Collins, M. (2008). Simple Semi-supervised Dependency Parsing. In *Proceedings of ACL*.

Liakata, M. and Pulman, S. (2008). Automatic Fine-Grained Semantic Classification for Domain Adaption. In *Proceedings of Semantics in Text Processing (STEP)*.

McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39.

McCarthy, J. (1986). Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 26(3):89–116.

Paşca, M. (2008). Turning Web Text and Search Queries into Factual Knowledge: Hierarchical Class Attribute Extraction. In *Proceedings of AAAI*.

Pantel, P. and Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of KDD*.

Schubert, L. K. (2002). Can we derive general world knowledge from texts? In *Proceedings of HLT*.

Singh, P. (2002). The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*. AAAI.

Steyvers, M. and Griffiths, T. (2007). Probabilistic Topic Models. In Landauer, T. K., Kintsch, W., McNamara, D. S., and Dennis, S., editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Inc.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of WWW*.

Van Durme, B., Michalak, P., and Schubert, L. K. (2009). Deriving Generalized Knowledge from Corpora using WordNet Abstraction. In *Proceedings of EACL*.