Knowledge Supervision as a Grounding for Machine Learning

by Mike Bergman - Tuesday, June 23, 2015

http://www.mkbergman.com/1872/knowledge-supervision-as-a-grounding-for-machine-learning/

Providing the Method behind Knowledge-based Artificial Intelligence

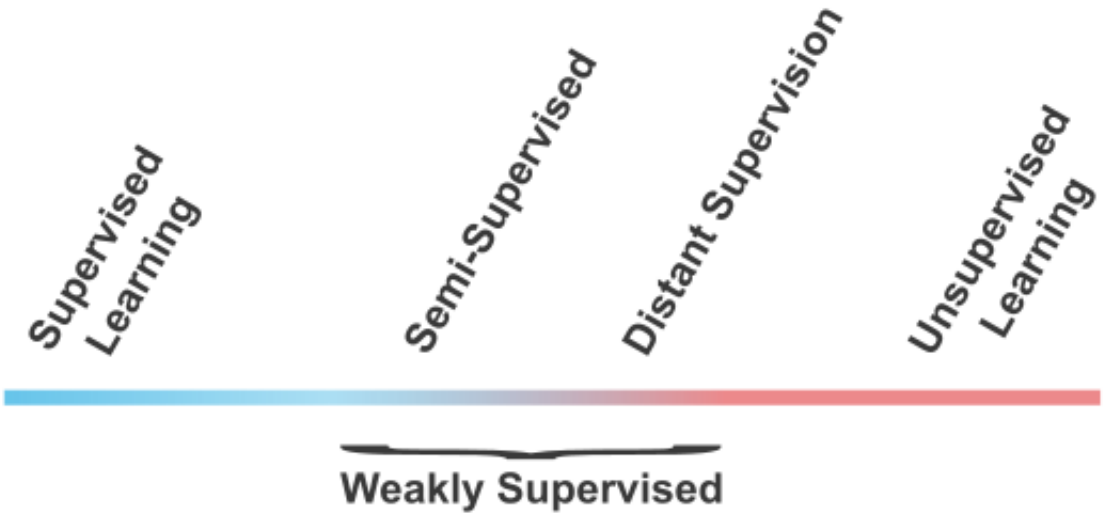
One of the central engines behind artificial intelligence is machine learning. ML involves various ways that data is used to train or teach machines to classify, predict or perform complicated tasks, such as I captured in an earlier diagram. The methods used in machine learning may be statistical, based on rules, or recognizing or discovering patterns.

The name machine learning begs the question of *to learn what*? In the context of images, audio, video or sensory perception, machine learning is trained for the recognition of patterns, which can be layered into learning manifolds called deep learning. In my realm -- that is, knowledge bases and semantics -- machine learning can be applied to topic or entity clustering or classification; entity, attribute or relation identification and extraction; disambiguation; mapping and linking multiple sources; language translation; duplicates removal; reasoning; semantic relatedness; phrase identification; recommendation systems; and, question answering. Significant results can be obtained in these areas without the need for deep learning, though that can and is being usefully applied in areas like machine translation or artificial writing.

Machine learning can be either supervised or unsupervised. In supervised learning, positive and (often) negative training examples are presented to the learning algorithm in order to create a model to produce the desired results for the given context. No training examples are presented in unsupervised learning; rather, the model is derived from patterns discovered in the absence of training examples, sometimes described as finding hidden patterns in unlabeled data. Supervised methods are generally more accurate than unsupervised methods, and nearly universally so in the realm of content information and knowledge.

There is effort and expense associated with creating positive or negative training examples (sets). This effort can span from the maximum of ones completely constructed manually to ones that are semi-automatic (semi-supervised) or to ones informed by knowledge bases (weakly supervised or distant supervised [1], [2]). Creation of manual training sets may consume as much as 80% of overall efforts in some cases, and is always a time-consuming task whenever employed. The accuracy of the eventual models is only as good as the trueness of the input training sets, with traditionally the best results coming

from manually determined training sets; the best of those are known as "gold standards." The field of machine learning is thus broad and multiple methods span these spectra of effort and accuracy.



The Spectrum of Machine Learning

To date, the state-of-the-art in machine learning for natural language processing and semantics, my realm, has been in distant supervision using knowledge bases like Freebase or Wikipedia to extract training sets for supervised learning [1]. Relatively clean positive and negative training sets may be created with much reduced effort over manually created ones. This is the current "sweet spot" in the accuracy *v.* effort trade-off for machine learning in my realm.

However, as employed to date, distant supervision has mostly been a case-by-case, problem-by-problem approach, and most often applied to entity or relation extraction. Yes, knowledge bases may be inspected and manipulated to create the positive and negative training examples needed, but this effort has heretofore not been systematic in approach nor purposefully applied across a range of ML applications. How to structure and use knowledge bases across a range of machine learning applications with maximum accuracy and minimum effort, what we call *knowledge supervision*, is the focus of this article. The methods of *knowledge supervision* are how we make operational the objectives of knowledge-based artificial intelligence. This article is thus one of the foundations to my recent series on KBAI [3].

Features and Training Sets

Features and training sets, in relation to the specific machine learning approaches that are applied, are the major determinants to how successful the learning actually is. We already touched upon the trade-offs in effort and accuracy associated with training sets, and will provide further detail on this question below. But features also pose trade-offs and require similar skill in selection and use.

In machine learning, a feature is a measurable property of the system being analyzed. A feature is equivalent to what is known as an explanatory variable in statistics. A feature, stated another way, is a

measurable aspect of the system that provides predictive value for the state the system.

Features with high explanatory power independent of other features are favored, because each added feature adds a computational cost of some manner. Many features are correlated with one another; in these cases it is helpful to find the strongest signals and exclude the other correlates. Too many features also make tuning and refinement more difficult, what has sometimes been called the curse of dimensionality. Overfitting is also often a problem, which limits the ability of the model to generalize to other data.

Yet too few features and there is inadequate explanatory power to achieve the analysis objectives.

Though it is hard to find discussion of best practices in feature extraction, striking this balance is an art [4]. Multiple learners might also be needed in order to capture the smallest, independent (non-correlated) feature set with the highest explanatory power [5].

When knowledge bases are used in distant supervision, only a portion of their structure or content is used as features. Still other distant supervision efforts may be geared to other needs and use a different set of features. Indeed, broadly considered, knowledge bases (potentially) have a rich diversity of possible features arising from:

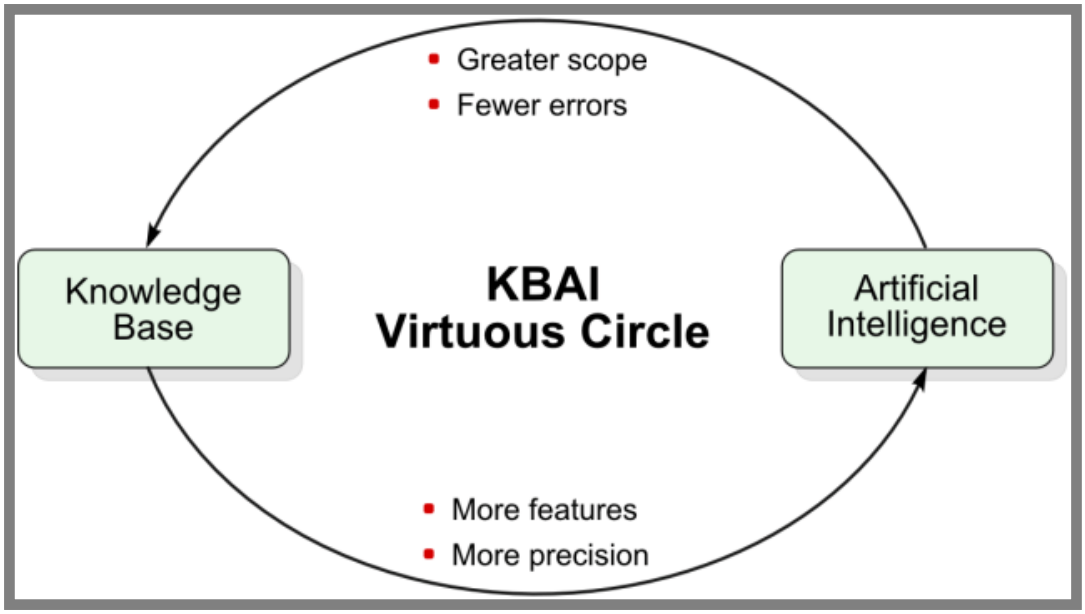
- text, and its content, syntax, semantics and morphology
- use vectors of co-occurring terms or concepts
- categories
- conventions
- synonyms
- linkages
- mappings
- relations
- attributes
- content placement within its knowledge graph, and
- disjointednesses.

An understanding of the features potential for knowledge bases is the first mindset of moving toward more purposeful knowledge supervision. At Structured Dynamics we stage the structured information as RDF triples and OWL ontologies, which we can select and manipulate via APIs and SPARQL. We also stage the graph structure and text in Lucene, which gives us powerful faceted search and other advanced NLP manipulations and analyses. These same features may also be utilized to extend the features set available from the knowledge base through such actions as new entity, attribute, or relation extractions; fine-grained entity typing [6]; creation of word vectors or tensors; results of graph analytics; forward or backward chaining; efficient processing structures; etc.

Because all features are selectable via either structured SPARQL query or faceted search, it is also possible to more automatically extract positive and negative training sets. Attention to proper coverage and testing of disjointedness assertions is another purposeful step useful to knowledge supervision, since it aids identification of negative examples for the training.

Whatever the combination of ML method, feature set, or positive or negative training sets, the ultimate precision and accuracy of knowledge supervision requires the utmost degree of true results in both positive and negative training sets. Training to inaccurate information merely perpetuates inaccurate information. As anyone who has worked extensively with source knowledge bases such as Freebase, DBpedia or Wikipedia may attest, assignment errors and incomplete typing and characterizations are all too common. Further, none provide disjointedness assertions.

Thus, the system should be self-learning with results so characterized as to be fed automatically to further testing and refinement. As better quality and more features are added to the system, we produce what we have shown before [3], as the virtuous circle of KBAI:



Features and training sets may thus be based on the syntax, morphology, semantics (meaning of the data) or relationships (connections) of the source data in the knowledge base. Continuous testing and the application of the system's own machine learners creates a virtuous feedback where the accuracy of the overall system is constantly and incrementally improved.

Manifest Applications for Knowledge Supervision

The artificial intelligence applications to which knowledge supervision may be applied are manifest. Here is a brief listing of some of those areas as evidenced by distant supervision applied to machine learning in academic research, or others not yet exploited:

- entity identification (recognition) and extraction
- attribute identification and extraction ("slot filling")
- relation identification and extraction
- event identification and extraction
- entity classifiers
- phrase (n-gram) identification
- entity linkers
- mappers

- topic clusterers
- topic classifiers
- disambiguators
- duplicates removal
- semantic relatedness
- inference and reasoning
- sub-graph extraction
- ontology matchers
- ontology mappers
- sentiment analysis
- question answering
- recommendation systems
- language translation
- multi-language versions
- artificial writing, and
- ongoing knowledge base improvements and extensions.

These areas are listed in rough order from the simpler to the more complex analyses. Most distant supervision efforts to date have concentrated on information extraction, the first items shown on the list. But all of these are amenable to knowledge supervision with ML. Since 2009, many of the insights regarding these potentials have arisen from the Knowledge Base Population initiative of the Text Analysis Conference [7].

Mapping and linkage are essential areas on this list since they add greatly to the available feature set and provide the bases for greater information interoperability. This is the current emphasis of Structured Dynamics.

Knowledge Supervision is Purposeful and Systematic

Knowledge supervision is the purposeful structuring and use of knowledge bases to provide features and training sets for multiple kinds of machine learners, which in combination can be applied to multiple artificial intelligence outcomes. Knowledge supervision is the method by which knowledge-based artificial intelligence, or KBAI, is achieved.

None of this is free, of course. Much purposeful work is necessary to configure and stage the data structures and systems that support the broad application of knowledge supervision. And other questions and challenges related to KBAI also remain. Yet, as Pedro Domingos has stated [4]:

"And the organizations that make the most of machine learning are those that have in place an infrastructure that makes experimenting with many different learners, data sources and learning problems easy and efficient, and where there is a close collaboration between machine learning experts and application domain ones."

Having the mindset and applying the methods of knowledge supervision produces an efficient, repeatable, improvable infrastructure for active learning about the enterprise's information assets.

As noted, we are just at the beginnings of knowledge supervision, and best practices and guidelines are still in the formative stages. We also have open questions and challenges in how features can be effectively selected; how KB-trained classifiers can be adopted to the wild; how we can best select and combine existing machine learners to provide an ML infrastructure; where and how deep learning should most effectively be applied; and how other emerging insights in computational linguistics can be combined with knowledge supervision [8].

But we can already see that a purposeful mindset coupled with appropriate metadata and structured RDF data is a necessary grounding to the system. We can see broad patterns across the areas of information extraction involving concepts, entities, relations, attributes and events that can share infrastructure and methods. We realize that linkage and mapping are key enabling portions of the system. The need for continuous improvement and codification of self-learning are the means by which our systems will get more accurate.

So, with the what of knowledge-based artificial intelligence, we can now add some broad understandings of the how based on knowledge supervision. None of these ideas are unique or new unto themselves. But the central role of knowledge bases in KBAI and knowledge supervision represents an important advance of artificial intelligence to deal with real-world challenges in content and information.

---

[1] Distant supervision was earlier or alternatively called *self-supervision*, *indirect supervision* or *weakly-supervised*. It is a method to use knowledge bases to label entities automatically in text, which is then used to extract features and train a machine learning classifier. The knowledge bases provide coherent positive training examples and avoid the high cost and effort of manual labelling. The method is generally more effective than unsupervised learning, though with similar reduced upfront effort. Large knowledge bases such as Wikipedia or Freebase are often used as the KB basis.

The first acknowledged use of distant supervision was Craven and Kumlien in 1999 (Mark Craven and Johan Kumlien. 1999. “Constructing Biological Knowledge Bases by Extracting Information from Text Sources,” in *ISMB*, vol. 1999, pp. 77-86. 1999; source of weak supervision term.)); the first use of the formal term distant supervision was in Mintz et al. in 2009 (Mike Mintz, Steven Bills, Rion Snow, Dan Jurafsky, 2009. “Distant Supervision for Relation Extraction without Labeled Data,” in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 1003–1011, Suntec, Singapore, 2-7 August 2009). Since then, the field has been a very active area of research; see next reference.

[2] See M. K. Bergman, 2015. "Forty Seminal Distant Supervision Articles," from *AI3:::Adaptive Information* blog, November 17, 2014, as supplemented by [3].

[3] See M. K. Bergman, 2014. "Knowledge-based Artificial Intelligence," from *AI3:::Adaptive Information* blog, November 17, 2014.

[4] Pedro Domingos, 2012. "A Few Useful Things to Know About Machine Learning." *Communications of the ACM* 55, no. 10 (2012): 78-87.

[5] There is a rich literature providing guidance on feature selection and feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the available

features. It is also possible to apply methods, the best known and simplest being principal component analysis, among many, to reduce feature size (dimensionality) with acceptable loss in accuracy.

[6] As a good introduction and overview, see Xiao Ling and Daniel S. Weld, 2012. "Fine-Grained Entity Recognition," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012*. You can also search on the topic in Google Scholar.

[7] TAC is organized by the National Institute of Standards and Technology (NIST). Initiated in 2008, TAC grew out of NIST's Document Understanding Conference (DUC) for text summarization, and the Question Answering Track of the Text Retrieval Conference (TREC). TAC is overseen by representatives from government, industry, and academia. The Knowledge Base Population tracks of TAC were started in 2009 and continue to today.

[8] See, for example, Percy Liang and Christopher Potts, 2015. "Bringing Machine Learning and Compositional Semantics Together." *Annu. Rev. Linguist.* 1, no. 1 (2015): 355-376.
