Data-Driven Computational Pragmatics

Class 1: Overview Notes

Pragmatics:

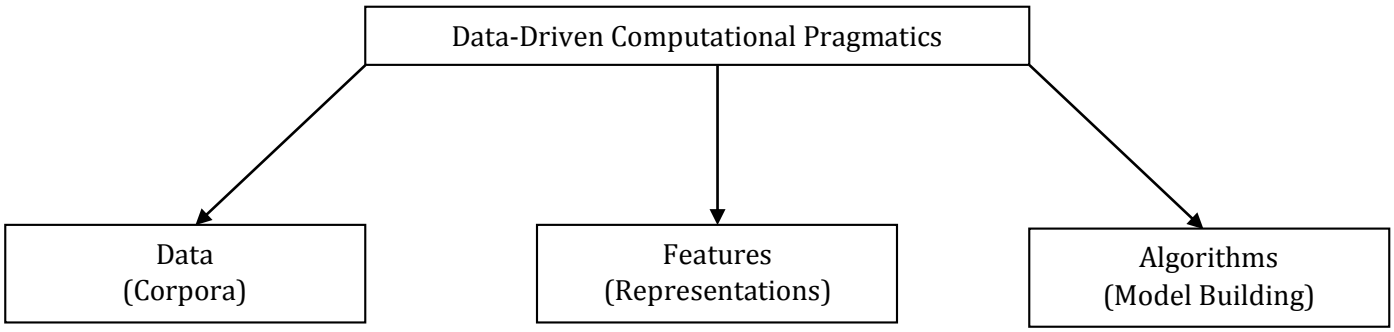
Meaning-in-Language has several sources, most prominently cognitive, contextual, and social meaning. Although no clear division between semantics and pragmatics is possible, pragmatics proto-typically focuses on contextual and social meaning. According to Searle, all three sorts of meaning are epistemically-objective in that they can be studied as naturally occurring phenomena and predictions can be made about them. They are ontologically-subjective, however, in that they depend upon human consciousness and are only directly accessible through human introspection and intuitions.

Computational:

Traditional linguistic analyses in semantics/pragmatics rely on introspection about meaning, intuition-based generation of counter-factuals, and models which make manual predictions (i.e., the relation between the model and its predictions is not direct). Computational linguistic analyses, on the other hand, build models which make direct predictions (i.e., for each input and each model there is a single prediction without human intervention) but do so without direct access to human introspections about meaning. Thus, the models themselves are entirely syntactic in nature.

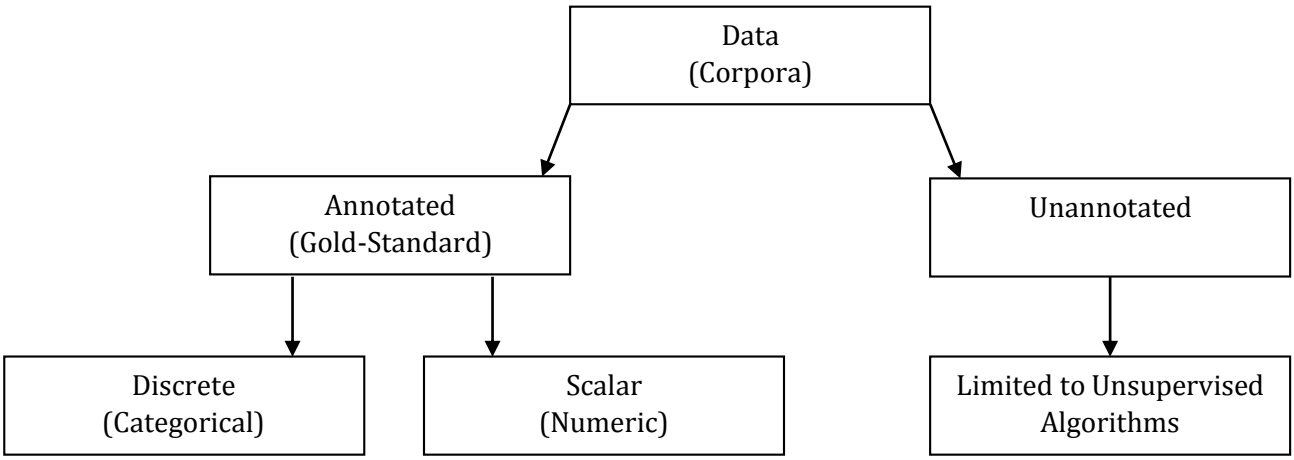
Data-Driven:

Computational models can be rule-based (e.g., formed by a set of heuristic, analyst created operations) or data-driven (e.g., derived from a dataset by a model-building algorithm and a set of features which represent the linguistic data). Data-driven models attempt to learn or derive a model directly from the linguistic data without a set of rule-based heuristics. This requires taking a corpus-based view of language in which language is an ontologically-objective produced phenomenon (i.e., with no access to introspections about meaning). Most often, machine learning is used to produce or learn models from the data. Model-building algorithms cannot directly interact with language data, so language must be represented as features extracted from the corpus. Feature extraction is essentially automated annotation.



Data:

Data, in the form of corpora, is essential for data-driven approaches. Corpora are large collections of language, usually as text. Often, but not always, corpora have been collected as a representative sample of a particular group. Size and representativeness of the dataset used are important restrictions on the results of individual studies.



Annotated Data (Gold-Standard):

Corpora are often annotated for particular linguistic and non-linguistic attributes. Gold-standard annotations refer to manual annotations of the phenomenon addressed by a particular model. Thus, gold-standard annotations are the ideal or correct output or predictions of the model. Gold-standard annotations are usually done manually by a number of annotators with reported inter-annotator agreement. Those annotations which are used by the model to make predictions, usually computationally derived, are called features (see next page).

Categorical vs. Scalar Annotations:

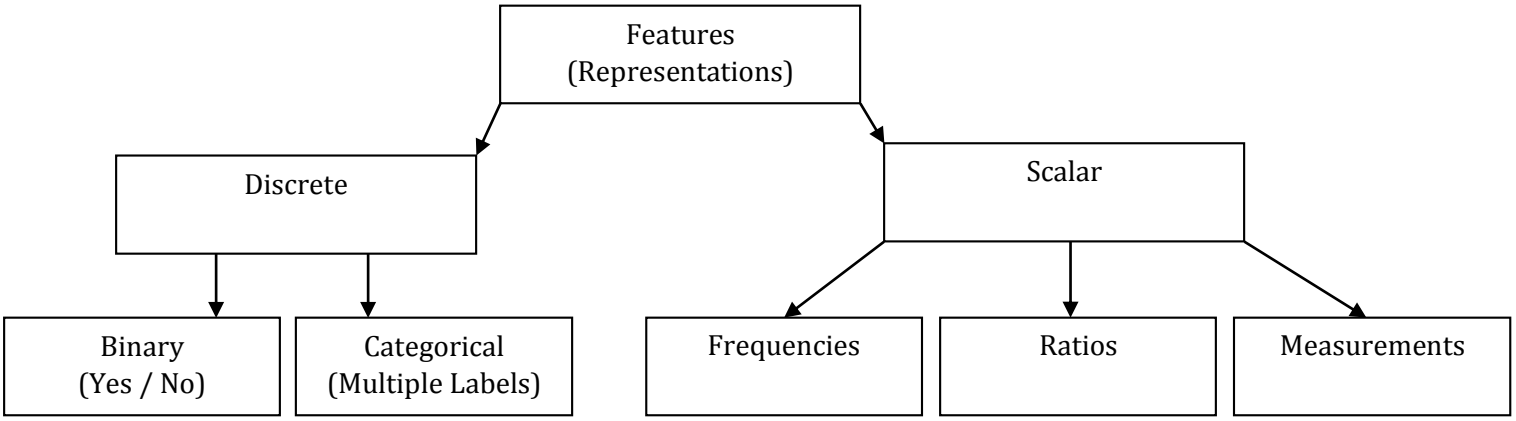
Annotations, like any variable, can be categorical (e.g., discrete) or scalar (e.g., numeric). For example, categorical annotations can be multi-valued labels (e.g., an entity could be a PERSON, PLACE, ORGANIZATION, etc. in entity recognition) or can be binary (e.g., two nouns could be co-referencing or not co-referencing in many models of co-reference). Numeric annotations, on the other hand, refer to a particular scale: examples include abstractness or metaphoricity, both derived from psycholinguistic participant-based studies. Categorical annotations can be modeled using a classifier while scalar attributes, unless discretized, must be modeled using a regression.

Unannotated Data:

Datasets without gold-standard annotations are limited to unsupervised model-building algorithms, or clustering algorithms. Models sometimes have multiple levels: for example, some co-reference models first classify individuals pairs of nouns as co-referenced or not, and then cluster co-referenced nouns into chains which reference the same entity.

Features (Representations):

Features are computationally derived annotations or representations of language data used as variables in a model. The type of features used in a particular model constitute a hypothesis about what properties of language are relevant for predicting the phenomenon in question. Many model-building algorithms allow redundant and unnecessary features (i.e., they are simply not included in the final model, or given limited weight). Thus, it is customary to test various sub-groups of features in order to see which contribute to the final predictions and to what extent.



Discrete Features:

Discrete representations can be binary or categorical. Binary features often represent whether a given unit of language satisfies a given condition (for example, whether or not a sentence is in the present tense, whether or not a noun carries gender information). Thus, these binary features often are used to represent rule-based heuristics even within a data-driven model. Categorical features (for example, the semantic role of a given noun), likewise, depend upon a feature extraction component which might itself rely on rule-based heuristics. Because models often contain other models as sub-parts (e.g., having a semantic parser as a dependency), there is not always a sharp distinction between data-driven and rule-based models.

Scalar Features: Frequency-based:

Numeric representations of language have several different sources. Most commonly these features are frequency counts of a particular property within a given unit of language (e.g., word n-grams in a text). Frequency counts are usually represented as raw counts, as relative frequency counts, or scaled using something like the Term-Frequency / Inverted-Document-Frequency (TF-IDF) measure. Scaled features like the TF-IDF represent a form of learning in themselves; thus, for example, they are heavily dependent on the reference corpus used to scale the measure.

Scalar Features: Ratio-based:

Ratio-based scalar features are those which balance one scalar variable in relation to another (e.g., the proportion of nouns to adjectives in a sentence or the proportion of lexical words to function

words). Ratio features can also be used to generate complex features from basic features; for example, entropy-guided feature induction first generates many combinations of features (e.g., Feature A * Feature B) and then uses an information gain measure to evaluate the generated features according to their usefulness in a particular model.

Scalar Features: Measurement-based:

Measurement-based features are those which reflect a measurement of some property of the language. For example, word-level abstractness ratings are often used as lexical features. These abstractness ratings are determined using participant-based psycholinguistic studies and assign each word to a scale between abstract and concrete words. The feature thus represents a psycholinguistic measurement, contained in a dictionary, which is used to represent individual words. In this way, outside measurements and psycholinguistic findings can be incorporated into the model.

Level of Representation:

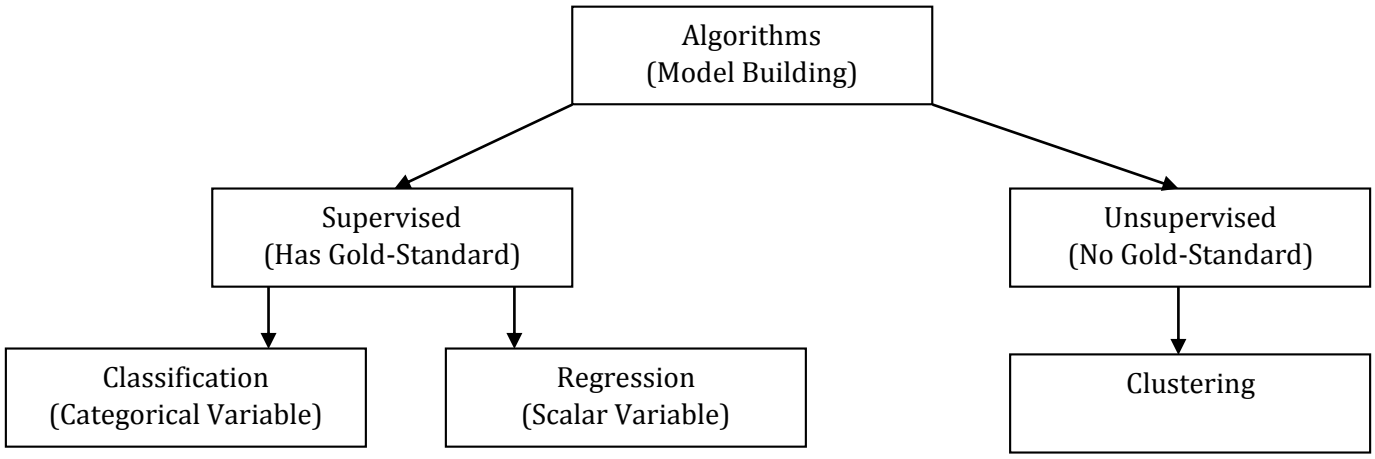
Features thus can represent language at multiple levels of representation. Many features are word-based (e.g., n-gram frequency features). Others are based on properties of clauses or sentences (e.g., a binary feature indicating whether a sentence is passive or active). Still other features operate at the level of whole texts. Thus, the features which a model uses hypothesize that the predicted phenomenon acts at a particular level of representation.

Feature Extraction:

Data-driven models use computationally derived features. Thus, the step of feature extraction or automatic representation of language data is an essential component of such models. At the same time, data-driven models often contain rule-based or heuristic-based feature extraction components which are not themselves data-driven.

Algorithms (Model-building):

Given a set of features to represent language and a set of gold-standard annotations describing the phenomenon in question, learning algorithms are used to build a model of the phenomenon. The model uses the features to make predictions about the phenomenon.



Supervised Algorithms:

Supervised learning algorithms depend on a class variable from gold-standard annotation. These algorithms attempt to learn generalizations about the relationship between the class variable and the features representing the language data in a training corpus, build a model of this relationship, and then use the features to make predictions on new test data. Supervised algorithms require curating a set of annotated training data which is large enough, representative enough, and balanced enough to learn generalizable models.

Classification:

Supervised learning algorithms with a categorical class variable are called classifiers. Prominent classifiers include Naïve Bayes, Logistic Regression, Support Vector Machines, and J48.

Regression:

Supervised learning algorithms with a scalar class variable are based on regression analysis. A prominent example is Linear Regression.

Unsupervised Algorithms: Clustering

Unsupervised learning algorithms do not have gold-standard class variables and thus do not require training data. Rather, the algorithms use features to cluster units of language (ranging from co-reference pairs to whole texts) into groups of related objects. Thus, clustering algorithms can be thought of as assigning class membership (e.g., cluster membership) to data without class data, thus creating classes based on the features. Common unsupervised clustering algorithms include K-Means (a centroid-based model of clusters), Expectation Maximization (a distribution-based model of clusters), and DBSCAN (a density-based model of clusters).