Abstract

In most of the languages today, it has been observed that the sequence of words follow a well-defined structure. This underlying structure is basically the syntax of the language. Also, it is shown in previous researches that syntactic features are the most informative features in lexical class discovery such as verb classification. This work is motivated to enrich under-resourced languages such as Hindi with a well-developed lexicon. We want to design a framework that will be unsupervised and can induce lexical classes and syntax from the given data. We may further want to extend this framework for languages such as Hindi which is poor in terms of resources.

Related Work Done

Integrating Topics and Syntax

- Generative model that uses short-range and long-range dependencies to find syntactic and semantic classes simultaneously
- A composite model in which syntactic component is HMM and semantic component is LDA.
- The model is defined in terms of three sets of variables: words, topics, and classes.
- They have used Markov Chain Monte Carlo (MCMC) to perform full Bayes inference in this model, sampling from a posterior distribution over assignments of words to classes and topics.
- Brown and TASA corpora were used. The Brown corpora had 500 documents and 1,137,466 word tokens, with POS tags for each token.
- The TASA corpora was untagged and had 37,561 documents along with 12,190,931 word tokens.

Syntactic Topic Models

- A non-parametric Bayesian Model of parsed documents.
- STM can generate words that are both thematically and syntactically constrained.
- The syntactic information is available from parse trees.
- STM is closely related to HDP and to infinite tree with independent children.
- Use variational inference rather than EM to compute posterior distribution.
- Latent variables of the model are topics, topic transition vectors, topic weights, topic assignments, and top-level weights.
- They have used synthetic data and Penn Tree Bank and compared with HDP.
- Except preposition and determiners, all other things were sub-grouped into themes.

Decision Tree for labeling Parts-of-Speech

- Uses decision trees for estimating marginal probabilities in a maximum entropy model given the context in which a word appears.
- Claim improvements over usual Hidden Markov Model approach.
- The algorithm for growing trees has as its objective function the entropy of the joint distribution of tags and histories. They minimize Kullback Information.
- Mutual information clustering is used to obtain a set of binary features to assign to words.
- Created a model for tagging text using a portion of Lancaster Treebank and Tagged a portion of Brown corpus.
- Tag vocabulary size was 163 tags and word vocabulary size was 41471 words.
- The HMM errors was 3.03% whereas for DT was 2.61%.

POS induction through Prototype Discovery

- Fully unsupervised algorithm inspired by cognitive notion of prototypes.
- Algorithm first identifies landmark cluster of words, which are essentially the prototypes, and then maps rest of the words to these clusters.
- Work is evaluated on English and German
  1) First cluster words based on fine morphological representation.
  2) Cluster most frequent words, define landmark clusters.
- Maps the rest of the words to these clusters.
- For distributional representation, left and right context is used and for clustering, they use average-link clustering.
- For English the model was trained on 39832 sentences from section 2-21 of PTB-WSJ and on 500K sentences from NYT section of NANC newswire corpus.
- There are 45 clusters, 34 of which are not punctuation. Used k=14 and k=34.

Word Representations

- To improve existing systems for NER and Chunking by inserting unsupervised word representations as extra word features into them.
- Evaluated Brown clusters, Collobert and Weston (2008) embeddings, and HLBL embeddings (Mnih & Hinton, 2009) and various combinations of word representations.
- Classified word representations into Distributional, Cluster-based and Distributed.
- Used unsupervised data for inducing word representations.
- The data for chunking is from Penn Treebank, and is newswire from Wall Street Journal in 1989.
- The standard evaluation benchmark for NER is the CoNLL03 shared task dataset drawn from the Reuters newswire.
- Brown clusters perform best for both NER and chunking.

NLP from Scratch

- Proposes a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including: part-of-speech tagging, chunking, named entity recognition, and semantic role labelling.
- Do not exploit man-made input features carefully optimized for each task.
- System learns internal representations on the basis of vast amounts of mostly unlabelled training data.
- Basis for building a freely available tagging system with excellent performance while requiring minimal computational resources.

To Do

- Running the existing code on Hindi data
- Improving the implementation of ADIOS
- Merging lexical class induction with ADIOS to extract syntactic and lexical classes

References