Application of Bayesian Framework in Natural Language Understanding

Authors: Pawan Goyal, Laxmidhar Behera, T.M. McGinnity

Presented by: Burcu Can
Contents

1. Introduction
2. Basic Theory
3. Inference in Bayesian Networks
4. Extensions of Bayesian Networks
5. Bayesian Networks and Natural Language Understanding
   - Part-of-Speech Tagging
   - Parsing of a Sentence
   - Word Sense Disambiguation
   - Machine Translation
6. Summary
Introduction

• Problems in NLP:
  – Data sparseness
  – Spelling variants/errors (‘airplane’, ‘aeroplane’ or ‘foetus’, ‘fetus’)
  – Ambiguity (‘saw’ – a tool or the past tense of the verb ‘see’)
  – Pronoun resolution
Introduction – ‘cont

• Techniques using machine learning
  – State machines
  – Neural networks
  – Genetic algorithms etc.

• Nowadays, the dominant approach
  – Bayesian networks
Introduction – ‘cont

• Reasons for using Bayesian Networks:
  • Extension of probabilistic models
  • Explicitly represent the conditional dependencies
  • Provides an intuitive graphical visualization of the knowledge
  • Representation of conditional independence assumptions
  • Representation of the joint probability distribution of the model.
    • Less probabilities of the probabilistic model
    • Reduced computational complexity of the inferences
Basic Theory

- S-Snow, CL-Clouds, R-Rain, F-Flood, A-Car accident in a street, T-Traffic Jam, D-Delay, C-Causality
Basic Theory – ‘cont

• Term similarity between Traffic Jam (T) and Rain (R):

  \[ \text{term-sim}(T,R) = P(T|R) + P(R|T) \]
  \[ = P(T|R) + P(T|R)\frac{P(R)}{P(T)} \]
  \[ = P(T|A)P(A|R)(1 + \frac{P(R)}{P(T)}) \]

  which is an example of inferencing in Bayesian network.
Inference in Bayesian Networks

- $P(X_i \mid E)$
  - $E$: set of evidence variables
  - Decompose $E$ into two parts:
    - $E^-$ is the part consisting of assignments to variables in the subtree rooted at $X_i$
    - $E^+$ is the rest of it.
- $P(X_i \mid E) = P(X_i \mid E^-, E^+) = P(E^- \mid X_i, E^+) P(X \mid X_i, E^+) / P(E^- \mid E^+) = \alpha \lambda (X_i) \pi (X_i)$
  - $\lambda (X_i) = P(E^- \mid X_i)$
  - $\pi (X_i) = P(X_i \mid E^+)$
Inference in Bayesian Networks

- Ex: Cloud(CL) and Delay(D) are observed in the document, and inference on Car Accident(A) is wanted.
  - $P(X_i | E)$?
    - $X_i : A$
    - $E : CL, D, E^- = CL, E^+ = D$
    - $P(A | CL, D) = \alpha \lambda(A) \pi(A)$
    - $\alpha = 1/P(D | CL)$
    - $\lambda = P(CL | A)$
    - $\pi = P(A | D)$
Extensions of Bayesian Networks

- Dynamic Bayesian Networks (DBN)
- Hierarchical Bayesian Networks
- Sigmoid Bayesian Networks
- Incremental Sigmoid Belief Networks (ISBN)
- Mixed Bayesian Networks

- Others: Belief networks, inference networks
Bayesian Networks and Natural Language Understanding

- Part-of-Speech (POS) Tagging
- Word Sense Disambiguation
- Machine Translation
- Information Retrieval
Part-of-Speech Tagging

• The process of marking up the words based on its definition, as well as its context:
  – nouns, adjectives, adverbs etc.

• Ex: The sailor dogs the hatch.
Part-of-Speech Tagging - 'cont

• Peshkin uses DBN for POS tagging.
• Forward-backward algorithm is used for the inference.
• POS tags are viewed as time series data of the observed samples in a given sentence.
Part-of-Speech Tagging - 'cont

Feature set:
- Capitalization
- Hyphenation
- Numeric
- Prefix
- Suffix

Observable features
- Memory

Prefix
- Suffix
- Word
- Number
- Hyphen
- Case

Index $N$

Index $N+1$
Part-of-Speech Tagging - 'cont

• The probability of a complete sequence of POS tags $T_1...T_n$ is modeled as:

$$
\Pr(T_1...T_n) = \Pr(T_1) \times \Pr(F_1 | F_1) \times \Pr(T_2 | T_1, \text{Start}) \\
\times \Pr(F_2 | T_2) \times \Pr(M_2 | T_1) \\
\times \prod_{i=3}^{n-1} \Pr(T_i | T_{i-1}, M_{i-1}) \\
\times \Pr(M_i | T_{i-1}, M_{i-1}) \times \Pr(F_i | T_i) \\
\times \Pr(T_n | T_{n-1}, M_{n-1}) \times \Pr(F_n | T_n),
$$

$$
\Pr(F_i | T_i) = \Pr(S_i | T_i) \times \Pr(P_i | T_i) \times \Pr(W_i | T_i) \\
\times \Pr(C_i | T_i) \times \Pr(H_i | T_i) \times \Pr(N_i | T_i)
$$
Word Sense Disambiguation (WSD)

• Task of finding the sense of a word in a context.
  – Ex: the word *bass*
    • 1. a type of fish
    • 2. tones of low frequency

• 1. I went fishing for some sea bass
• 2. The bass line of the song is too weak
Word Sense Disambiguation (WSD) - 'cont

• Necessary in:
  – Machine Translation
  – Information Retrieval
  – Information Extraction

• Two Bayesian methods, naïve Bayes classifier and Bayesian belief networks are used to learn probabilistic classifiers for WSD.
Word Sense Disambiguation (WSD) - 'cont

• Bayesian Belief Networks [Wiebe, Bruce]
  – Community
    • People living in a particular area
    • An association of people with similar interests
    • Common ownership
    • The body of people in a learned occupation
  – Town
    • An urban area with a fixed boundary that is smaller than a city
    • The people living in a municipality smaller than a city
    • An administrative division of a county
Word Sense Disambiguation (WSD) - 'cont

• The one node per sense approach
Word Sense Disambiguation (WSD) - 'cont
• The one node per word approach
Word Sense Disambiguation (WSD) - 'cont

• Synset: a group of data elements that are considered semantically equivalent for the purposes of information retrieval.

• Hypernym: conceptual parent

• Ex: synset: \{occupation, vocation, occupational group\} is the hypernym of synset: \{profession, community\}
Word Sense Disambiguation (WSD) - 'cont

• Defining the CPTs:
  – Each child node corresponds to a hyponym node in WordNet.
  – Assign conditional probability:
    • $P(\text{hyponym} | \text{hypernym})$
    • Ex: MUNICIPALITY#1 has two children

<table>
<thead>
<tr>
<th>municipality#1</th>
<th>P(town#1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.000 + \epsilon</td>
</tr>
<tr>
<td>T</td>
<td>0.500</td>
</tr>
</tbody>
</table>
Machine Translation

- The task of translating the text from one natural language to another.
- Static Bayesian networks, dynamic Bayesian networks
- Filali has introduced a new generalization of DBN, as multi dynamic Bayesian networks (MDBN)
- MDBN has multiple streams of variables that can get unrolled, but where each stream may be unrolled for a differing amount.
Machine Translation – ‘cont

•MDBN is a variant of DBN.
•DBN consists of a directed acyclic graph
  \[ G = (V, E) = (V_1 \cup V_2, E_1 \cup E_2 \cup E_2^\rightarrow) \]
Machine Translation – ‘cont

• Multi-Dynamic Bayesian Network (MDBN)

\[ G = (V, E) = \left( \bigcup_{k} V^{(k)}, \bigcup_{k} E^{(k)} \cup E'_{\parallel} \right) \]

• IBM Model
Machine Translation – ‘cont

• A string of French words $F=f$ of length $M=m$, into a string of English words $E=e$ of length $L=l$.
  – $P(f,e) = P(f|e)P(e)$, $P(e)$: language model
  – $P(f,e) = \sum_a P(f,a|e)$
Information Retrieval

• Extracting useful information from document collections

• Document retrieval inference network
  – Document network
  – Query network
Information Retrieval – ‘cont

• Document network
  – Document nodes \((d_i)\)
  – Text representation nodes \((t_j)\)
  – Concept representation nodes \((r_k)\)

• Query network
  – Information need \((I)\)
  – Query nodes \((q_i)\)
  – Query concept nodes \((c_i)\)
Information Retrieval – ‘cont
Information Retrieval – ‘cont

• Belief Network Model (Ribeiro-Neto)

- $d_i$: document
- $k_t$: index terms
- $q$: user query
- $d = k_1, k_2, \ldots, k_t$
- $q = k_1, k_2, \ldots, k_t$
Information Retrieval – ‘cont

• \( P(d \mid q) = ? \) (ranking of document \( d \), with respect to the query \( q \))

• \( P(q) = \sum_u P(q \mid u) \cdot P(u) \) query concept

• \( P(d) = \sum_u P(d \mid u) \cdot P(u) \)

• \( P(d \mid q) = \frac{P(d \land q)}{P(q)} \Rightarrow P(d \mid q) \propto P(d \land q) \)

• \( P(d \mid q) \propto \sum_u P(d \land q \mid u) P(u) \)

• \( P(d \mid q) \propto \sum_u P(d \mid u) P(q \mid u) P(u) \)
Information Retrieval – ‘cont

• Bayesian Network Retrieval Model (Campos et al.)
  – Flexible topology
  – Term relationships
  – Document relationships
  – Fast, exact inference
  – No query components, instead query is considered as an evidence
Information Retrieval – ‘cont

- Bayesian Network Retrieval Model
  - Two sets of variables:
    - Terms ($\alpha=T_i, i=1,...,M$)
    - Documents ($\beta=D_j, j=1,...,N$)
  - Two types of knowledge
    - Expert knowledge
    - Collection knowledge
Information Retrieval – ‘cont

• Bayesian Network Retrieval Model
  – Expert knowledge
    • Connection of term nodes and documents
    • No links joining the documents
  – Collection knowledge
    • Dependence relationships between the terms
    • Uses polytree
Information Retrieval – ‘cont

• Bayesian Network Retrieval Model
Information Retrieval – ‘cont

- Fung et. al
  - Topic relationships
Information Retrieval – ‘cont

• Semantic Bayesian Networks (Hong et al)
  – Keyword layer
  – Concept layer
  – Target layer
Information Retrieval – ‘cont
Information Retrieval – ‘cont

- How to assign conditional probabilities?

| Condition                                      | $P(w = 1 | c_i = 1)$ | $P(w = 1 | c_i = 0)$ |
|------------------------------------------------|----------------------|----------------------|
| $w$ must occur given $c_i$                     | 0.95-0.99            | 0.7-0.9              |
| $w$ often occurs given $c_i$                  | 0.95-0.99            | 0.2-0.5              |
| $w$ may occur given $c_i$                     | 0.4-0.6              | 0.01-0.1             |
| $w$ seldom occurs given $c_i$                 | 0.2-0.3              | 0.01-0.1             |
| $w$ never occurs given $c_i$                  | 0.01-0.1             | 0.01-0.1             |
Information Retrieval – ‘cont

• Semantic relations

<table>
<thead>
<tr>
<th>Class</th>
<th>Sub-classification</th>
<th>Relationship</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has-a</td>
<td>Object-attribute</td>
<td>O – A</td>
<td>Phone-bell, MP3 player-price</td>
</tr>
<tr>
<td></td>
<td>Attribute-value</td>
<td>A – V</td>
<td>Size-big, price-low</td>
</tr>
<tr>
<td>Is-a</td>
<td>-</td>
<td>Is-a</td>
<td>Size-volume</td>
</tr>
</tbody>
</table>
Information Retrieval – ‘cont

- Sride et. al
  - Information Retrieval System for Structured Documents based on Bayesian networks
Information Retrieval – ‘cont

• Structured document representation:
Information Retrieval – ‘cont

• Structured document representation:
Summary & Discussion

- Four NLP problems are given.
- Can cope up with well with NLP problems
- Could be combined with other ML methods
References

• Brown, Cocke et al, A statistical Approach to Machine Translation, 1990
• Campos et al, Applying Bayesian Networks to Information Retrieval, 1995
• Peshkin et al, Part-of-Speech Tagging with minimal lexicalization, 2003
• Peshkin et al, Bayesian Nets for Syntactic Categorization of Novel Words, 2003
• Janyce et al, Constructing Bayesian Networks from WordNet for Word-Sense Disambiguation, 1998
• Kim et al, A Semantic Bayesian Network Approach to Retrieving Information with Intelligent Conversational Agents
• Turtle et al, Inference Networks for Document Retrieval
• Ribeiro-Neto et al, A Belief Network Model for Information Retrieval, 1996
Questions?