Machine Learning Approaches for Natural Language Processing

WANG Haifeng (王海峰)
Toshiba (China) R&D Center
MLA, Nov. 7, 2009
Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT
## CL vs. NLP

<table>
<thead>
<tr>
<th>Computational Linguistics, CL</th>
<th>Natural Language Processing, NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACL:</strong> Association for Computational Linguistics</td>
<td><strong>EMNLP:</strong> Empirical Methods in Natural Language Processing</td>
</tr>
<tr>
<td><strong>COLING:</strong> International Conference on Computational Linguistics</td>
<td><strong>IJCNLP:</strong> International Joint Conference on Natural Language Processing</td>
</tr>
<tr>
<td><strong>ICCL:</strong> International Committee on Computational Linguistics</td>
<td><strong>AFNLP:</strong> Asian Federation of Natural Language Processing</td>
</tr>
<tr>
<td><strong>CNCCL:</strong> Chinese National Conference on Computational Linguistics</td>
<td><strong>YSSNLP:</strong> Young Scholar Symposium on Natural Language Processing</td>
</tr>
<tr>
<td><strong>ICL:</strong> Institute of Computational Linguistics</td>
<td><strong>NLPLAB:</strong> Natural Language Processing LAB</td>
</tr>
</tbody>
</table>

**Impact?**

**History**

**Theory**

**Methodology**
## ACL-IJCNLP 2009

<table>
<thead>
<tr>
<th>Area</th>
<th>#Submission</th>
<th>#Accepted</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation</td>
<td>82</td>
<td>23</td>
<td>28.0%</td>
</tr>
<tr>
<td>Semantics</td>
<td>67</td>
<td>14</td>
<td>20.9%</td>
</tr>
<tr>
<td>Syntax and Parsing</td>
<td>49</td>
<td>14</td>
<td>28.6%</td>
</tr>
<tr>
<td>Information Extraction</td>
<td>49</td>
<td>10</td>
<td>20.4%</td>
</tr>
<tr>
<td>Discourse, Dialogue and Pragmatics</td>
<td>43</td>
<td>9</td>
<td>20.9%</td>
</tr>
<tr>
<td>Summarization and Generation</td>
<td>44</td>
<td>8</td>
<td>18.2%</td>
</tr>
<tr>
<td>Phonology, Morphology, Segmentation, POS, Chunking</td>
<td>31</td>
<td>8</td>
<td>25.8%</td>
</tr>
<tr>
<td>Sentiment Analysis, Opinion Mining, Classification</td>
<td>45</td>
<td>7</td>
<td>15.6%</td>
</tr>
<tr>
<td>Statistical and Machine Learning Methods</td>
<td>40</td>
<td>6</td>
<td>15.0%</td>
</tr>
<tr>
<td>Spoken Language Processing</td>
<td>19</td>
<td>6</td>
<td>31.6%</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>28</td>
<td>4</td>
<td>14.3%</td>
</tr>
<tr>
<td>Language Resource</td>
<td>26</td>
<td>4</td>
<td>15.4%</td>
</tr>
<tr>
<td>Text Mining and NLP Applications</td>
<td>21</td>
<td>4</td>
<td>19.0%</td>
</tr>
<tr>
<td>Question Answering</td>
<td>25</td>
<td>3</td>
<td>12.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>569</strong></td>
<td><strong>120</strong></td>
<td><strong>21.1%</strong></td>
</tr>
</tbody>
</table>
## NLP Areas

### ACL-IJCNLP 2009

<table>
<thead>
<tr>
<th>Area</th>
<th>#Submission</th>
<th>#Accepted</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation</td>
<td>82</td>
<td>23</td>
<td>28.0%</td>
</tr>
<tr>
<td>Semantics</td>
<td>67</td>
<td>14</td>
<td>20.9%</td>
</tr>
<tr>
<td>Syntax and Parsing</td>
<td>49</td>
<td>14</td>
<td>28.6%</td>
</tr>
<tr>
<td>Information Extraction</td>
<td>49</td>
<td>10</td>
<td>20.4%</td>
</tr>
<tr>
<td>Discourse, Dialogue and Pragmatics</td>
<td>43</td>
<td>9</td>
<td>20.9%</td>
</tr>
<tr>
<td>Summarization and Generation</td>
<td>44</td>
<td>8</td>
<td>18.2%</td>
</tr>
<tr>
<td>Phonology, Morphology, Segmentation, POS, Chunking</td>
<td>31</td>
<td>8</td>
<td>25.8%</td>
</tr>
<tr>
<td>Sentiment Analysis, Opinion Mining, Classification</td>
<td>45</td>
<td>7</td>
<td>15.6%</td>
</tr>
<tr>
<td>Statistical and Machine Learning Methods</td>
<td>40</td>
<td>6</td>
<td>15.0%</td>
</tr>
<tr>
<td>Spoken Language Processing</td>
<td>19</td>
<td>6</td>
<td>31.6%</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>28</td>
<td>4</td>
<td>14.3%</td>
</tr>
<tr>
<td>Language Resource</td>
<td>26</td>
<td>4</td>
<td>15.4%</td>
</tr>
<tr>
<td>Text Mining and NLP Applications</td>
<td>21</td>
<td>4</td>
<td>19.0%</td>
</tr>
<tr>
<td>Question Answering</td>
<td>25</td>
<td>3</td>
<td>12.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>569</strong></td>
<td><strong>120</strong></td>
<td><strong>21.1%</strong></td>
</tr>
</tbody>
</table>
NLP Taxonomy

- **Sub-task**
  - Analysis & understanding, generation

- **Level**
  - Morphology, syntax, semantics, pragmatics

- **Grammar**
  - PS, DS, LFG, HPSG, CCG …

- **Unit**
  - Character, word, phrase, sentence, paragraph …

- **Style**
  - Spoken language, written language

- **Application**
  - Translation, information retrieval and extraction, sentiment, QA, summarization, grammar check …

- **Approach**
  - Rationalist and empiricist approaches

- **Data**
  - Lexicon, rules, corpus (labeled and unlabeled)
Difficulties

- Complex structure
  - Mapping between string and structure

- Ambiguities
  - Disambiguation

- Examples
  - 打：打酱油、打毛衣、打人、打针 ……
  - pretty little girls' school
    - Does the school look little?
    - Do the girls look little?
    - Do the girls look pretty?
    - Does the school look pretty?
Approaches

- Rationalist approaches
  - Linguistic theory
  - Grammar system
  - Rules
    - Usually manually compiled
  - Popular in NLP application (e.g. RBMT)

  It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

- Empiricist approaches
  - Corpus
    - Labeled, unlabeled
    - Monolingual, multilingual
  - Statistical and Machine Learning Approaches
  - Dominant approach in NLP research

  Whenever I fire a linguist our system performance improves.

Noam Chomsky

Frederick Jelinek
Data vs. Algorithm

Data

Algorithm
Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT
Why ML in NLP

- Practical concerns
  - Data
    - More and more unlabeled language data available on the web and elsewhere
    - Labeled data are easier to create than rules
  - Method
    - It is hard for knowledge engineering methods to cope with the growing flood of data
    - Machine learning can be used to automate knowledge acquisition and inference
  - Computing resource
    - More and more powerful (Moore’s law)

- Theoretical contribution
  - Reasonably solid foundations (theory and algorithms)

ML gives elegant, well-founded solutions to NLP problems

NLP comes with data and gives meaning to ML's math
Designing ML for NLP

- Data
  - How to access and use data

- Target function
  - Concept to be learnt

- Representation
  - Representation of hypotheses
  - Representation of data

- Learning algorithm
  - Conditioned to the representation
  - Inductive learning assumption
ML applications in NLP

NLP Areas
- Machine Translation
- Semantics
- Syntax and Parsing
- Information Extraction
- Information Retrieval
- Summarization and Generation
- Question Answering
- ……
- ML methods have been used in most NLP areas

ML Methods
- HMM, ME, CRF, SVM, Boosting, Co-training ……
- Many ML methods have been or will be used in NLP
ML at Leading NLP Conferences

ACL-IJCNLP 2009

- 2 sessions on “Statistical and Machine Learning Methods” (1 best paper)
  - Improving learning method
  - Mapping Instructions to Actions (one of the best papers)
  - Semantic
  - Word Segmentation
  - POS

- Much more ML related papers in other sessions
  - Invited talk
    - Heterogeneous Transfer Learning with Real-world Applications (Qiang Yang)
  - Machine translation
  - Parsing
  - Semantics
  - Question Answering
  - Information Extraction
  - ......
Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT
Machine Translation

To translate text from one natural language to another by computer

English, Chinese, Arabic, Japanese, Korean, European languages, etc.

Machine Translation

Computer Science
Linguistics
Cognitive Science
Information Science

Knowledge Engineering
Software Engineering
## Applications

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information distribution</td>
<td>Word</td>
</tr>
<tr>
<td>Information gathering</td>
<td>Phrase</td>
</tr>
<tr>
<td>Communication</td>
<td>Sentence</td>
</tr>
<tr>
<td>CLIR</td>
<td>Paragraph</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Translation Mode</th>
<th>Product Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic MT</td>
<td>Software package</td>
</tr>
<tr>
<td>Computer Aided Translation</td>
<td>MT engine license</td>
</tr>
<tr>
<td>Speech Translation</td>
<td>Translation service</td>
</tr>
<tr>
<td></td>
<td>Hardware preinstall</td>
</tr>
</tbody>
</table>
How

Rule based MT

Rationalist

Rules

Dictionary

Empiricist

Example based MT
Statistical MT

Corpus
Methods

RBMT
- Good at describing linguistic phenomena
- Dominant method in commercial product
- Potential for Improvement (Zhu. 2005)

EBMT
- Good at translating similar sentence
- Still popular (MT Journal, MT Summit)
- Comparable with SMT (Way 2005, Liu 2006)

SMT
- Model, learning, robustness
- Dominant in MT research
- Won NIST MT evaluations
Combination

RBMT

SMT

EBMT

Wu, Chiang, Groves, Liu
He’s just taken the medicine.

- Hierarchical
- Fine grained
- Scalable
Example-based machine translation (EBMT)
- Machine translation by example-guided inference, or machine translation by the analogy principle (Nagao, 1984)

Three main components
- Match fragments against a database of real examples
- Identify the corresponding translation fragments
- Combine these to give the final translation

Performance
- Good performance in domain specific application
Input: The city is China’s principal capital market.
Overview of SMT

- Word-based SMT (Brown et al., 1990 & 1993)
- Phrase-based SMT (Koehn et al., 2003)
- Syntax-based SMT (Wu, 1997; Chiang, 2005)
**Method**

- **Statistical theory**
  \[
  \hat{e}_i^l = \arg \max_{e_i^l} \{ \Pr(e_i^l | f_i^J) \}
  \]

- **Generative method**
  - Translation process is broken down into steps
  - Each step is modeled by a probability distribution
  - Each probability distribution is estimated by maximum likelihood

- **Discriminative method**
  - Model consists of a number of features
  - Each feature has a weight
  - Feature weights are optimized on development set
Source-channel model

\[ \hat{e}_1^I = \arg \max_{e'_1^I} \{ \Pr(e_1^I) \cdot \Pr(f_1^J | e_1^I) \} \]

Log-linear model

\[ \hat{e}_1^I = \arg \max_{e'_1^I} \left\{ \frac{\exp \left[ \sum_{m=1}^{M} \lambda_m h_m (e_1^I, f_1^J) \right]}{\sum_{e''_1^I} \exp \left[ \sum_{m=1}^{M} \lambda_m h_m (e''_1^I, f_1^J) \right]} \right\} \]

\[ = \arg \max_{e_1^I} \{ \exp \left[ \sum_{m=1}^{M} \lambda_m h_m (e_1^I, f_1^J) \right] \} \]
Word-based SMT – IBM Model 1

- Only uses lexical translation
- Lexical translation probabilities is estimated from a parallel corpus
- Chicken and egg problem
  - if we had the alignments,
    - we could estimate the parameters of our generative model
  - if we had the parameters,
    - we could estimate the alignments
- EM algorithm
  - Initialize model parameters (e.g. uniform)
  - Assign probabilities to the missing data
  - Estimate model parameters from completed data
  - Iterate
## Word-based SMT – Higher IBM Models

<table>
<thead>
<tr>
<th>IBM Model 1</th>
<th>Lexical translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 2</td>
<td>Adds absolute reordering model</td>
</tr>
<tr>
<td>IBM Model 3</td>
<td>Adds fertility model</td>
</tr>
<tr>
<td>IBM Model 4</td>
<td>Relative reordering model</td>
</tr>
<tr>
<td>IBM Model 5</td>
<td>Fixes deficiency</td>
</tr>
</tbody>
</table>

- Training of a higher IBM model builds on previous model
Word-based SMT – IBM Model 3

From (Knight and Koehn)
Phrase-based SMT

**<Training>**

*Parallel corpus*

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'm from England.</td>
<td>I am English.</td>
</tr>
<tr>
<td>Lots of people enjoy drinking Chinese tea.</td>
<td>大勢の人々は中国の茶を飲むのが好きです。</td>
</tr>
<tr>
<td>He loves reading.</td>
<td>彼は本を読むのが好きです。</td>
</tr>
</tbody>
</table>

*Automatic learning*

*Statistical models*

- **Translation model**
  - I → 私 (-3.4)
- **Language model**
  - 私はイギリス人です (-7.1)
  - は:イギリス → 人 (-5.5)

**<Translation>**

*Input: I love Chinese tea.*

**<Decoding>**

- 私はイギリス人です
- が好き
- 中国の
- 好き
- です

*Translation: 私は中国のお茶が好きです.*

- Foreign input is segmented in phrases
  - Not necessarily linguistically motivated
- Each foreign phrase is translated into native phrase
  - Search a phrase table
- Phrases are reordered
Syntax-based SMT

- Why syntax in SMT
  - More grammatical output
  - Syntax aware re-ordering
  - Accurate insertion of function words

- Grammars
  - Synchronous Context Free Grammars (SCFG)
  - Linguistically informed grammars

- Models
  - Tree-to-String
  - String-to-Tree
  - Tree-to-Tree

- Disadvantage
More Data, Better Translations

From (Koehn, 2003: Europarl Corpus)
Outline

- Overview of Natural Language Processing (NLP)
- Machine Learning Approaches for NLP
- Overview of Machine Translation (MT)
- Semi-Supervised Boosting for Statistical Word Alignment and SMT
Motivation

- Using both labeled and unlabeled data
- Unlabeled data
  - Aligned automatically
  - IBM models
  - Large, easy
- Labeled data
  - Aligned by human
  - Following a given alignment standard
  - Small, hard
- Adjust automatic alignment results by using manually aligned data
Boosting

Initialization -> Learner

Learner -> Calculate Error Rate

Calculate Error Rate -> Re-weight Training data

Re-weight Training data -> Reference Set

Reference Set -> End?

End? -> Yes

Yes -> Build Ensemble
Semi-Supervised Boosting

- Three main problems
  - Semi-supervised learner
    - Combine labeled data and unlabeled data
  - Reference set
    - Automatically construct a reference set for unlabeled data
  - Error rate calculation
    - How to calculate the error rate with both labeled data and unlabeled data
Semi-Supervised Boosting Word Alignment

1. **Labeled Data**
   - Supervised Training

2. **Unlabeled Data**
   - Unsupervised Training

3. **Model Interpolation**

4. **Real Reference Set**
   - Error Rate Calculation

5. **Re-weight Training data**

6. **End?**
   - Yes

7. **Pseudo Reference Set**
   - Ensemble
Word Alignment Model

- Supervised alignment model
  - Calculate the probabilities for IBM Model based on the labeled data

- Unsupervised alignment model
  - EM training for IBM Model

- Perform model interpolation

\[
\Pr(\mathbf{a}, \mathbf{f} | e) = \lambda \cdot \Pr_S(\mathbf{a}, \mathbf{f} | e) + (1 - \lambda) \cdot \Pr_U(\mathbf{a}, \mathbf{f} | e)
\]
Pseudo Reference Set Construction

- Obtain bi-directional word alignment sets $S_1$ and $S_2$ on the training data
- Obtain the intersection set of these two alignment sets
  \[ I = S_1 \cap S_2 \]
- Filter the union set of the two alignment sets
  \[ C = \{(s,t) \mid p(t \mid s) > \delta_1 \& count(s,t) > \delta_2 \} \]
  where
  \[ p(t \mid s) = \frac{\text{count}(s,t)}{\sum_{t} \text{count}(s,t)} \]
- Build the pseudo reference set
  \[ R = I \cup C \]
Error Rate Calculation

- For a sentence pair

\[AER(i) = 1 - \frac{2 \cdot |S_G \cap R_S|}{|S_G| + |R_S|}\]

- Calculate the error rate of a aligner
  - Based on the labeled data instead of the whole data

\[\varepsilon_l = \sum_{i \in D} w_l(i) \cdot AER(i)\]

where

\[w_l(i)\] is the normalized weight of the \(i^{th}\) sentence pair at the \(l^{th}\) round
Re-weight the Training Data

- Reweight each sentence pair in the training set
  - For each sentence pair, there may exists correct links and incorrect links as compared with the pseudo reference set
  - Calculate the weight of each sentence pair according to the correct and incorrect links

\[
w_{l+1}(i) = w_l(i) \times \left(k + (n - k) \times \beta_l\right) / n
\]

where

\[
\beta_l = \varepsilon_l / (1 - \varepsilon_l)
\]

- \(K\) is the number of the error links
- \(n\) is the total number of the links in the reference
Final Ensemble

- Obtain the final ensemble according to the trained word aligners on each round

\[ h_f(s) = \arg \max_t \sum_{l=1}^{L} (\log \frac{1}{\beta_l}) \cdot WT_l(s,t) \cdot \delta(h_l(s) = t) \]

where \( WT_l(s,t) = \frac{2 \cdot \text{count}(s,t)}{\sum_{t'} \text{count}(s,t') + \sum_{s'} \text{count}(s',t)} \)

- \( WT_l(s,t) \) is the weight of each alignment pair (s,t) produced by the word aligner \( h_l \)
- \( h_f \) is the final ensemble for word alignment
- \( \log \frac{1}{\beta_l} \) is the weight of the word aligner \( h_l \)
Evaluation

- **Training set**
  - Unlabeled data: 320,000 English-Chinese pairs
  - Labeled data: 30,000 English-Chinese pairs

- **Held-out set**
  - 1,500 sentence pairs

- **Testing set**
  - 1,000 bilingual English-Chinese sentence pairs
  - Totally 8,651 alignment links
  - 866 multi-word alignment links
Evaluation Metric

- **Word alignment**
  - Precision and Recall
  - Alignment Error Rate (AER)

\[
\text{Precision} = \frac{|S_G \cap S_C|}{|S_G|} \quad \text{Recall} = \frac{|S_G \cap S_C|}{|S_C|}
\]

\[
\text{AER} = 1 - \frac{2|S_G \cap S_C|}{|S_G| + |S_C|}
\]

- **Phrase-based machine translation**
  - BLEU, NIST
Word Alignment Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpolation</td>
<td>0.7555</td>
<td>0.7084</td>
<td>0.2688</td>
</tr>
<tr>
<td>Supervised Boosting</td>
<td>0.7771</td>
<td>0.6757</td>
<td>0.2771</td>
</tr>
<tr>
<td>Unsupervised Boosting</td>
<td>0.8056</td>
<td>0.7070</td>
<td>0.2469</td>
</tr>
<tr>
<td>Semi-supervised Boosting</td>
<td>0.8175</td>
<td>0.7858</td>
<td>0.1987</td>
</tr>
</tbody>
</table>
Weights in Ensembles

- Two kinds of weights
  - Weights for the individual aligners
  - Weights for the individual alignment links

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.7946</td>
<td>0.7775</td>
<td>0.2140</td>
</tr>
<tr>
<td>Our method</td>
<td>0.8175</td>
<td>0.7858</td>
<td>0.1987</td>
</tr>
</tbody>
</table>

Baseline: only use the first kind of weights
Our method: use the two kinds of weights
## Translation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpolation</td>
<td>4.7929</td>
<td>0.1350</td>
</tr>
<tr>
<td>Supervised Boosting</td>
<td>4.4296</td>
<td>0.1151</td>
</tr>
<tr>
<td>Unsupervised Boosting</td>
<td>4.9045</td>
<td>0.1459</td>
</tr>
<tr>
<td>Semi-supervised Boosting</td>
<td>5.1729</td>
<td>0.1525</td>
</tr>
</tbody>
</table>
DEMO