Semi-supervised methods of text processing, and an application to medical concept extraction

Yacine Jernite
Text-as-Data series
September 17. 2015
What do we want from text?

1. Extract information
2. Link to other knowledge sources
3. Use knowledge (Wikipedia, UpToDate, …)
How do we answer those questions?

1. What do people talk about on social media, and how? (Sentiment analysis)

2. What actions are described in a news article? (Semantic parsing)

3. In a medical setting: what symptoms does a patient exhibit?
Pipeline
Pipeline

Text

Information, Features
Pipeline

Low level NLP

Text

IE algorithms

Information, Features
Pipeline

Low level NLP

IE algorithms

End-to-end NLP

Information, Features
Machine learning approach

1. Specify task
2. Specify training algorithm
3. Get data
4. Train
Machine learning approach

1. Specify task
2. Specify training algorithm
3. Get data
4. Train
So much text, so few labels

- 5M English Wikipedia articles (3G words)
- 54M Reddit comments
- 1G Words in Gigaword dataset (newswire text)
- 5-grams from 1T words
So much text, so few labels

- 1M words in Penn TreeBank (parsing)
- Machine translation: highly language (and domain) dependent
- A few thousand to few hundred thousand sentence…
- And so many other custom tasks
1. Literature review on semi-supervised paradigms
   a. Label induction
   b. Feature learning

2. Current work: Semi-Supervised Medical Entity Linking
## Overview

### Label induction

1. Labeling data is costly
2. Automatically obtain approximate labeling on larger dataset
3. Train using pseudo-labels
Overview

Feature learning

1. Feature quality affects accuracy
2. Learn features using other sources
3. Train with features on small labeled dataset
So much text, so few labels

- **Label induction**
- **Feature learning**
- **Domain adaptation**
- **Multi-view learning**
Overview

Labels

• Fine Grained Entity Recognition
  • Ling and Weld, 2012
• Distant Supervision for RE with an incomplete KB
  • Min et al., 2013
• Co-Training for DA
  • Chen et al. 2011
• Semi-Supervised FSP for Unknown Predicates
  • Das and Smith, 2011
Fine Grained Entity Recognition

- **Method type:** Automatic labeling
- **Task:** Identify entities in text, and tag them with one of 112 types
- **Labeled data:** Hand-labelled news reports
- **Auxiliary data:** Wikipedia, Freebase
Fine Grained Entity Recognition

Freebase

Don Quixote

Don Quixote, fully titled The Ingenious Gentlemen Don Quixote of La Mancha, is a Spanish novel by Miguel de Cervantes Saavedra. It follows the adventures of Alonso Quijano, an hidalgo who reads so many chivalric novels that he loses his sanity and decides to set out to revive chivalry, undo wrongs, and bring justice to the world, under the name Don Quixote. He recruits a simple farmer, Sancho Panza, as his squire, who often employs a unique, earthy wit in dealing with Don Quixote's rhetorical orations on antiquated knighthood. Don Quixote, in the first part of the book, does not see the world for what it is, and prefers to imagine that he is living out a knighthood story. The story implements various themes, such as intertextuality, realism, metatheatre, and literary representation. Published in two volumes, in 1605 and 1615, Don Quixote is considered the most influential work of literature from the Spanish Golden Age and the entire Spanish literary canon.
Fine Grained Entity Recognition

1. Automatically label entity spans in Wikipedia text

Don Quixote

... 

Meaning

Harold Bloom says that *Don Quixote* is the writing of radical nihilism and anarchy,...
Fine Grained Entity Recognition

1. Automatically label Wikipedia text
   ● Spans are obtained from hyperlinks
   ● Types are obtained from Freebase

Don Quixote

... Meaning
Harold Bloom says that Don Quixote is the writing of radical nihilism and anarchy,…

Harold Bloom: Topic, Academic, Person, Author, Award winner, Influence node

Nihilism: Topic, Field of study, Literature subject, Religion
Fine Grained Entity Recognition

1. Train CRF and perceptron on pseudo-labeled data

Harold Bloom says that Don Quixote is a person book
Fine Grained Entity Recognition

- Compares to
  - Stanford NER: 4 most common classes
  - Ratinov et al. Named Entity Linking

- Results:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Strict</th>
<th>Loose Macro</th>
<th>Loose Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEL</td>
<td>0.220</td>
<td>0.327</td>
<td>0.381</td>
</tr>
<tr>
<td>Stanford (CoNLL)</td>
<td>0.425</td>
<td>0.585</td>
<td>0.548</td>
</tr>
<tr>
<td>FIGER</td>
<td>0.471</td>
<td>0.617</td>
<td>0.597</td>
</tr>
<tr>
<td>FIGER (GOLD)</td>
<td>0.532</td>
<td>0.699</td>
<td>0.693</td>
</tr>
</tbody>
</table>
Distant Supervision for Relation Extraction with an incomplete Knowledge Base

- **Method type**: Automatic labeling, Label inference
- **Task**: Relation extraction
- **Labeled data**: TAC 2011 KBP dataset
- **Auxiliary data**: Wikipedia infoboxes, Freebase
Distant Supervision for Relation Extraction with an incomplete Knowledge Base

- Entity pairs extracted from Wikipedia infoboxes
- Labeled with FreeBase relations: origin
Distant Supervision for Relation Extraction with an incomplete Knowledge Base

- Latent variable algorithm to learn from positive-only labels

- \( X \): entity pair mention
- \( Z \): mention level label
- \( \ell \): bag level label
- \( Y \): KB entity pair label
- \( \theta \): Number of positive labels
Distant Supervision for Relation Extraction with an incomplete Knowledge Base

- Learns with EM, compares to \((y = l)\)
Co-Training for Domain Adaptation

- **Method type**: Automatic labeling, Domain adaptation
- **Task**: Text classification - review polarity
- **Labeled data**: Amazon reviews for books, DVD, electronics, kitchen
- **Auxiliary data**: Cross-domain training
Self-Training

Labeled data

Unlabeled data
Self-Training

- Labeled data
- Classifier 1
- Unlabeled data
- Pseudo-labeled data
Self-Training

Labeled data

Classifier 1

Unlabeled data

Pseudo-labeled data

Classifier 2
Self-Training

● Algorithm
  ● Train System-1 on labeled data
  ● Label some data with System-1
  ● Train System-2 on combined data

● Not much improvement
  ● Less than 1% parsing accuracy
  ● Somewhat better “portability”
Co-Training

Labeled data

Unlabeled data
Co-Training

Labeled data → Classifier 1 → Classifier 2

Unlabeled data

Pseudo-labeled data
Co-Training

Labeled data → Classifier 1 → Classifier 2 → Selection

Unlabeled data → Pseudo-labeled data
Co-Training

Labeled data

Classifier 1

Classifier 2

Unlabeled data

Selection

Pseudo-labeled data
Co-Training

- Algorithm
  - Train System-1 and System-2 on labeled data with disjoint feature sets
  - Add data which is confidently labeled by exactly one system
  - Re-train, iterate

- Theoretical guarantees for “independent” feature sets
Co-Training for Domain Adaptation

- L1 regularization: starts using more target-domain features
Co-Training for Domain Adaptation

- Best improvement adding a limited number of examples
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- **Method type**: Label pre-selection
- **Task**: Frame-semantic parsing
- **Labeled data**: SemEval 2007
- **Auxiliary data**: Gigaword corpus, FrameNet
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

Ted really **tried** to read *Infinite Jest*, but was discouraged by the size of the book.

**Attempt**

**Definition:**

An **Agent** attempts to achieve a **Goal**. The **Outcome** may also be mentioned explicitly.

**John** **ATTEMPTED** to climb Mt. Everest.

It was another **failed** **ATTEMPT** to climb Mt. Everest.
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Extracts possible frame targets from unlabeled data
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Extracts possible frame targets from unlabeled data
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Graph construction
  - Distance from dependency parsed text
  - About 60,000 targets (about 10,000 in FrameNet)
  - Convex quadratic optimization problem
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

- Learned neighbor frame distribution

<table>
<thead>
<tr>
<th></th>
<th>$t = discrepancy.N$</th>
<th></th>
<th>$t = contribution.N$</th>
<th></th>
<th>$t = print.V$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>$q_t^*(f)$</td>
<td>$f$</td>
<td>$q_t^*(f)$</td>
<td>$f$</td>
<td>$q_t^*(f)$</td>
</tr>
<tr>
<td>*SIMILARITY</td>
<td>0.076</td>
<td>*GIVING</td>
<td>0.167</td>
<td>*TEXT_CREATION</td>
<td>0.081</td>
</tr>
<tr>
<td>NATURAL_FEATURES</td>
<td>0.066</td>
<td>MONEY</td>
<td>0.046</td>
<td>SENDING</td>
<td>0.054</td>
</tr>
<tr>
<td>PREVARICATION</td>
<td>0.012</td>
<td>COMMITMENT</td>
<td>0.046</td>
<td>DISPERASAL</td>
<td>0.054</td>
</tr>
<tr>
<td>QUARRELING</td>
<td>0.007</td>
<td>ASSISTANCE</td>
<td>0.040</td>
<td>READING</td>
<td>0.042</td>
</tr>
<tr>
<td>DUPLICATION</td>
<td>0.007</td>
<td>EARNINGS_AND_LOSSES</td>
<td>0.024</td>
<td>STATEMENT</td>
<td>0.028</td>
</tr>
</tbody>
</table>
Semi-Supervised Frame-Semantic Parsing for Unknown Predicates

• Parsing results

<table>
<thead>
<tr>
<th>Model</th>
<th>Unknown Targets</th>
<th>All Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact Match</td>
<td>Partial Match</td>
</tr>
<tr>
<td></td>
<td>$P$  $R$  $F_1$</td>
<td>$P$  $R$  $F_1$</td>
</tr>
<tr>
<td>SEMAFOR</td>
<td>19.59 16.48 17.90</td>
<td>33.03 27.80 30.19</td>
</tr>
<tr>
<td>Self-training</td>
<td>15.44 13.00 14.11</td>
<td>29.08 24.47 26.58</td>
</tr>
<tr>
<td>LinGraph</td>
<td>29.74 24.88 27.09</td>
<td>44.08 36.88 40.16</td>
</tr>
<tr>
<td>FullGraph</td>
<td>35.27* 28.84* 31.74*</td>
<td>48.81* 39.91* 43.92*</td>
</tr>
</tbody>
</table>
Overview

Features

- Prototype-Driven Learning for Sequence Models
  - Haghighi and Klein, 2006
- DA with Structural Correspondence Learning
  - Blitzer et al., 2006
- NLP (almost) from scratch
  - Collobert et al., 2011
- On Using Monolingual Corpora in NMT
  - Gulcehere et al., 2015
Prototype-Driven Learning for Sequence Models

- **Method type**: Feature learning
- **Task**: POS tagging, Classified ads segmentation
- **Labeled data**: PTB/CTB, Classifieds
- **Auxiliary data**: Prototypes
Prototype-Driven Learning for Sequence Models

- Example prototypes:

<table>
<thead>
<tr>
<th>Label</th>
<th>Prototype</th>
<th>Label</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOMATES</td>
<td>roommate respectful drama</td>
<td>NNS</td>
<td>years shares companies</td>
</tr>
<tr>
<td>RESTRICTIONS</td>
<td>pets smoking dog</td>
<td>VBG</td>
<td>including being according</td>
</tr>
<tr>
<td>UTILITIES</td>
<td>utilities pays electricity</td>
<td>-LRB-</td>
<td>-LRB- -LCB-</td>
</tr>
<tr>
<td>AVAILABLE</td>
<td>immediately begin cheaper</td>
<td>DT</td>
<td>the a A The</td>
</tr>
<tr>
<td>SIZE</td>
<td>2 br sq</td>
<td>WPS</td>
<td>whose</td>
</tr>
<tr>
<td>PHOTOS</td>
<td>pictures image link</td>
<td>FW</td>
<td>bono del kanji</td>
</tr>
<tr>
<td>RENT</td>
<td>$ month <em>number</em>15*1</td>
<td>RP</td>
<td>Up ON</td>
</tr>
<tr>
<td>CONTACT</td>
<td><em>phone</em> call <em>time</em></td>
<td>VBD</td>
<td>said was had</td>
</tr>
<tr>
<td>FEATURES</td>
<td>kitchen laundry parking</td>
<td>$</td>
<td>$ US$ C$</td>
</tr>
<tr>
<td>NEIGHBORHOOD</td>
<td>close near shopping</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>ADDRESS</td>
<td>address carlmont <em>ordinal</em>5</td>
<td>; . !</td>
<td>; ;</td>
</tr>
</tbody>
</table>

Prototype examples include:

- Roommates: roommate respectful drama
- Restrictions: pets smoking dog
- Utilities: utilities pays electricity
- Available: immediately begin cheaper
- Size: 2 br sq
- Photos: pictures image link
- Rent: $ month *number*15*1
Prototype-Driven Learning for Sequence Models

- Gives prototypes of tag-token pairs
- Compute a similarity measure on tokens
- Adds similarity to the prototypes as a feature
Prototype-Driven Learning for Sequence Models

Results:

<table>
<thead>
<tr>
<th>Setting</th>
<th>Num Tokens</th>
<th>Classifieds segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>48K</td>
<td>42.2</td>
</tr>
<tr>
<td>PROTO</td>
<td>61.9</td>
<td>68.8</td>
</tr>
<tr>
<td>PROTO+SIM</td>
<td>79.1</td>
<td>80.5</td>
</tr>
</tbody>
</table>

POS tagging

<table>
<thead>
<tr>
<th>Setting</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>46.4</td>
</tr>
<tr>
<td>PROTO</td>
<td>53.7</td>
</tr>
<tr>
<td>PROTO+SIM</td>
<td>71.5</td>
</tr>
<tr>
<td>PROTO+SIM+BOUND</td>
<td>74.1</td>
</tr>
</tbody>
</table>
Domain Adaptation with Structural Correspondence Learning

- **Method type:** Feature learning, Multi-view learning, Domain adaptation
- **Task:** POS tagging
- **Labeled data:** MEDLINE (target domain)
- **Auxiliary data:** WSJ (source domain)
Domain Adaptation with Structural Correspondence Learning

- Example: pivot features *required, from, for*

(a) An ambiguous instance

```
JJ vs. NN
with normal signal transduction
```

(b) MEDLINE occurrences of signal, together with pivot features

```
the signal required to stimulatory signal from essential signal for
```

(c) Corresponding WSJ words, together with pivot features

```
of investment required of buyouts from buyers to jail for violating
```
Domain Adaptation with Structural Correspondence Learning

● Defines a set of pivot features, present in both source and target

● Sets up a set of mini-tasks: “predict the presence of pivot feature f”

● Runs SVD on the learned weights $W_f$
Domain Adaptation with Structural Correspondence Learning

- Projection on first singular vector:
Domain Adaptation with Structural Correspondence Learning

- Results:

(b) Accuracy on 561-sentence test set

<table>
<thead>
<tr>
<th>Model</th>
<th>All</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratnaparkhi (1996)</td>
<td>87.2</td>
<td>65.2</td>
</tr>
<tr>
<td>supervised</td>
<td>87.9</td>
<td>68.4</td>
</tr>
<tr>
<td>semi-ASO</td>
<td>88.4</td>
<td>70.9</td>
</tr>
<tr>
<td>SCL</td>
<td>88.9</td>
<td>72.0</td>
</tr>
</tbody>
</table>

(c) Statistical Significance (McNemar’s) for all words

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>semi-ASO vs. super</td>
<td>0.0015</td>
</tr>
<tr>
<td>SCL vs. super</td>
<td>$2.1 \times 10^{-12}$</td>
</tr>
<tr>
<td>SCL vs. semi-ASO</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
NLP (almost) from Scratch

- **Method type**: Feature learning, Multi-view learning
- **Task**: POS, chunking, NER, SRL
- **Labeled data**: PTB, CoNLL
- **Auxiliary data**: 852M words from Wikipedia + Reuters
NLP (almost) from Scratch

- Neural network architecture
NLP (almost) from Scratch

- First approach: supervised training of neural networks for tasks

<table>
<thead>
<tr>
<th>Approach</th>
<th>POS (PWA)</th>
<th>Chunking (F1)</th>
<th>NER (F1)</th>
<th>SRL (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark Systems</td>
<td>97.24</td>
<td>94.29</td>
<td>89.31</td>
<td>77.92</td>
</tr>
<tr>
<td>NN+WLL</td>
<td>96.31</td>
<td>89.13</td>
<td>79.53</td>
<td>55.40</td>
</tr>
<tr>
<td>NN+SLL</td>
<td>96.37</td>
<td>90.33</td>
<td>81.47</td>
<td>70.99</td>
</tr>
</tbody>
</table>
Second approach: initialize with word representations from LM

<table>
<thead>
<tr>
<th>Approach</th>
<th>POS  (PWA)</th>
<th>CHUNK (F1)</th>
<th>NER  (F1)</th>
<th>SRL  (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark Systems</td>
<td>97.24</td>
<td>94.29</td>
<td>89.31</td>
<td>77.92</td>
</tr>
<tr>
<td>NN+WLL</td>
<td>96.31</td>
<td>89.13</td>
<td>79.53</td>
<td>55.40</td>
</tr>
<tr>
<td>NN+SLL</td>
<td>96.37</td>
<td>90.33</td>
<td>81.47</td>
<td>70.99</td>
</tr>
<tr>
<td>NN+WLL+LM1</td>
<td>97.05</td>
<td>91.91</td>
<td>85.68</td>
<td>58.18</td>
</tr>
<tr>
<td>NN+SLL+LM1</td>
<td>97.10</td>
<td>93.65</td>
<td>87.58</td>
<td>73.84</td>
</tr>
<tr>
<td>NN+WLL+LM2</td>
<td>97.14</td>
<td>92.04</td>
<td>86.96</td>
<td>58.34</td>
</tr>
<tr>
<td>NN+SLL+LM2</td>
<td>97.20</td>
<td>93.63</td>
<td>88.67</td>
<td>74.15</td>
</tr>
</tbody>
</table>
NLP (almost) from Scratch

- Finally: joint training

<table>
<thead>
<tr>
<th>Approach</th>
<th>POS (PWA)</th>
<th>CHUNK (F1)</th>
<th>NER (F1)</th>
<th>SRL (F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark Systems</td>
<td>97.24</td>
<td>94.29</td>
<td>89.31</td>
<td>77.92</td>
</tr>
<tr>
<td><strong>Window Approach</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN+SLL+LM2</td>
<td>97.20</td>
<td>93.63</td>
<td>88.67</td>
<td>–</td>
</tr>
<tr>
<td>NN+SLL+LM2+MTL</td>
<td>97.22</td>
<td>94.10</td>
<td>88.62</td>
<td>–</td>
</tr>
<tr>
<td><strong>Sentence Approach</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN+SLL+LM2</td>
<td>97.12</td>
<td>93.37</td>
<td>88.78</td>
<td>74.15</td>
</tr>
<tr>
<td>NN+SLL+LM2+MTL</td>
<td>97.22</td>
<td>93.75</td>
<td>88.27</td>
<td>74.29</td>
</tr>
</tbody>
</table>
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

- Sentiment analysis using word embeddings and syntactic parses
Skip-Thoughts Vectors
(Kiros et al., NIPS 2015)

- Encodes sentences directly
- Improves sentence-level tasks
  - Classification
  - Paraphrase
  - Image-sentence ranking
On Using Monolingual Corpora in NMT

- **Method type**: Feature learning, Target distribution
- **Task**: Machine Translation
- **Labeled data**: Aligned text
- **Auxiliary data**: Monolingual corpora
On Using Monolingual Corpora in NMT

- Neural Machine Translation as sequence to sequence modeling

- RNN encoder and decoder:
On Using Monolingual Corpora in NMT

- Train Neural Machine Translation system
- Train target language model: RNN
- Shallow fusion: beam search on combined scores
- Deep fusion: add language model hidden state as input to decoder (+controller)
On Using Monolingual Corpora in NMT

<table>
<thead>
<tr>
<th></th>
<th>tst2011</th>
<th>tst2012</th>
<th>tst2013</th>
<th>Test 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Best (Single)</td>
<td>18.77</td>
<td>18.62</td>
<td>18.88</td>
<td>-</td>
</tr>
<tr>
<td>Previous Best (Combination)</td>
<td>18.83</td>
<td>18.93</td>
<td>18.70</td>
<td>-</td>
</tr>
<tr>
<td>NMT</td>
<td>18.40</td>
<td>18.77</td>
<td>19.86</td>
<td>18.64</td>
</tr>
<tr>
<td>NMT+LM (Shallow)</td>
<td>18.48</td>
<td>18.80</td>
<td>19.87</td>
<td>18.66</td>
</tr>
<tr>
<td>NMT+LM (Deep)</td>
<td>20.17</td>
<td>20.23</td>
<td>21.34</td>
<td>20.56</td>
</tr>
</tbody>
</table>

Turkish
### On Using Monolingual Corpora in NMT

<table>
<thead>
<tr>
<th></th>
<th>SMS/CHAT</th>
<th></th>
<th>CTS</th>
<th></th>
<th>De-En</th>
<th></th>
<th>Cs-En</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>PB</td>
<td>15.5</td>
<td>14.73</td>
<td>21.94</td>
<td>21.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ CSLM</td>
<td>16.02</td>
<td>15.25</td>
<td>23.05</td>
<td>22.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPB</td>
<td>15.33</td>
<td>14.71</td>
<td>21.45</td>
<td>21.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ CSLM</td>
<td>15.93</td>
<td>15.8</td>
<td>22.61</td>
<td>22.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMT</td>
<td>17.32</td>
<td>17.36</td>
<td>23.4</td>
<td>23.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow</td>
<td>16.59</td>
<td>16.42</td>
<td>22.7</td>
<td>22.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep</td>
<td><strong>17.58</strong></td>
<td><strong>17.64</strong></td>
<td><strong>23.78</strong></td>
<td>23.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Chinese**
Semi-Supervised Learning for Entity Linkage using Variational Inference

Yacine Jernite, Alexander Rush and David Sontag
Semi-Supervised Learning for Entity Linkage using Variational Inference

- **Method type:** Feature learning, Label inference

- **Task:** Medical concept extraction

- **Labeled data:** Semeval 2015 (annotated medical notes)

- **Auxiliary data:** MIMIC-II (medical text), UMLS
Task description

- We have:
  - Medical text from the MIMIC database
  - Medical knowledge base UMLS with concept descriptions

- We want to identify concepts in the text and link them to UMLS
UMLS samples

- Ambiguous, incomplete

<table>
<thead>
<tr>
<th>C0027627</th>
<th>C0002895</th>
<th>C0342788</th>
</tr>
</thead>
<tbody>
<tr>
<td>neoplasm metastasis</td>
<td>anemia, sickle cell</td>
<td>renal carnitine transport defect</td>
</tr>
<tr>
<td>Neoplastic Process</td>
<td>Disease or Syndrome</td>
<td>Disease or Syndrome</td>
</tr>
<tr>
<td>metastases, neoplasm</td>
<td>transient abnormal myelopoiesis</td>
<td>carnitine uptake defect</td>
</tr>
<tr>
<td>metastasis</td>
<td>sickle cell anemia</td>
<td>systemic carnitine deficiency</td>
</tr>
<tr>
<td>secondaries</td>
<td>hemoglobin ss</td>
<td>scd</td>
</tr>
<tr>
<td>metastases</td>
<td>disease sickle-cell</td>
<td>primary carnitine defnecy</td>
</tr>
<tr>
<td>tumor cell migration</td>
<td>scd</td>
<td>cud</td>
</tr>
</tbody>
</table>
### UMLS samples

- Ambiguous, incomplete

<table>
<thead>
<tr>
<th>C0027627</th>
<th>C0002895</th>
<th>C0342788</th>
</tr>
</thead>
<tbody>
<tr>
<td>neoplasm metastasis</td>
<td>anemia, sickle cell</td>
<td>renal carnitine transport defect</td>
</tr>
<tr>
<td>Neoplastic Process</td>
<td>Disease or Syndrome</td>
<td>Disease or Syndrome</td>
</tr>
<tr>
<td>metastases, neoplasm metastasis</td>
<td>transient abnormal myelopoiesis</td>
<td>carnitine uptake defect</td>
</tr>
<tr>
<td>secondaries</td>
<td>sickle cell anemia</td>
<td>systemic carnitine deficiency</td>
</tr>
<tr>
<td>metastases</td>
<td>hemoglobin ss</td>
<td>primary carnitine deficiency</td>
</tr>
<tr>
<td>tumor cell migration</td>
<td>disease sickle-cell</td>
<td>cud</td>
</tr>
</tbody>
</table>

- scd
Step 1: Mention Detection

Head CT - "normal" per pt report, results not available - test was ordered for L sided hearing loss. Brief Hospital Course: 69yo M with HTN, Hyperlipidemia. Serial neurological exams did not reveal any new deficits.

Concept mention detection

Head CT - "normal" per pt report, results not available - test was ordered for \textit{L sided hearing loss}. Brief Hospital Course: 69yo M with \textit{HTN, Hyperlipidemia}. Serial \textit{neurological} exams did not reveal any new \textit{deficits}.
Step 1: Mention Detection

- B, I, O – ID, OD tagging with CRF

B I I I O

ordered for L sided hearing loss.

B OD OD OD OD OD ID

neuro exams did not reveal any deficits
Step 1: Mention Detection

- Duplicating incompatible examples

B I I I O O O
L sided hearing loss and pain.

B I OD OD OD ID O
L sided hearing loss and pain.
Step 1: Mention Detection

- Run inference on unlabeled and test set
- Approximate marginal probability
- Threshold
Step 1: Mention Detection

- PR curve:
Step 1: Mention Detection

- Other approaches:
  - ezDI: A Supervised NLP System for Clinical Narrative Analysis, Pathak et al., 2015
  - BIO for continuous, SVM to join
  - ULisboa: Recognition and Normalization of Medical Concepts, Leal et al., 2015
  - BIOENS tagging scheme, Brown clusters, domain lexicons
Step 2: Mention Identification

Head CT - "normal" per pt report, results not available - test was ordered for **L sided hearing loss**. Brief Hospital Course: 69yo M with **HTN, Hyperlipidemia**. Serial neurological exams did not reveal any new deficits.

<table>
<thead>
<tr>
<th>Concept</th>
<th>UMLS ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>L sided hearing loss</td>
<td>C0521785</td>
</tr>
<tr>
<td>HTN</td>
<td>C0020538</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>C0020473</td>
</tr>
<tr>
<td>Neurological deficits</td>
<td>CUI-less</td>
</tr>
</tbody>
</table>
Step 2: Mention Identification

- Pathak et al.:
  - Simple lookup
  - Semi-automated modified descriptions

- Edit distance

<table>
<thead>
<tr>
<th>CUI</th>
<th>Text</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0013132</td>
<td>Dribbling from mouth</td>
<td>Dribbling</td>
<td>from</td>
<td>mouth</td>
</tr>
<tr>
<td>C0014591</td>
<td>Bleeding from nose</td>
<td>Bleeding</td>
<td>from</td>
<td>nose</td>
</tr>
<tr>
<td>C0029163</td>
<td>Hemorrhage from mouth</td>
<td>Hemorrhage</td>
<td>from</td>
<td>mouth</td>
</tr>
<tr>
<td>C0392685</td>
<td>Chest pain at rest</td>
<td>Chest pain</td>
<td>at</td>
<td>rest</td>
</tr>
<tr>
<td>C0269678</td>
<td>Fatigue during pregnancy</td>
<td>Fatigue</td>
<td>during</td>
<td>pregnancy</td>
</tr>
</tbody>
</table>
Step 2: Mention Identification

- Leal et al.
  - Abbreviation dictionary
- UMLS lookup
- Similarity: Lucene, n-gram and edit distance
- Lowest Information Content (specificity, using UMLS tree structure)
Step 2: Mention Identification

- A Generative Entity-Mention Model for Linking Entities with KB (Han and Sun, ACL 2011)

- \[ p(m, e) = p(s, c, e) = p(e)p(s|e)p(c|e) \]

- \[ p(s|e) \]: translation model from main description

- \[ p(c|e) \]: unigram language model
Step 2: Mention Identification

- Our model:

\[ p(m, e) = p(m|e)p(e) \]

- \( p(m|e) \): multinomial with automatically curated support

- \( p(e) \): joint distribution on all entities in the document
Step 2: Mention Identification

- Our model:

\[ p(m, e) = p(m|e)p(e) \]

- \( p(m|e) \): multinomial with automatically curated support

- \( p(e) \): joint distribution on all entities in the document
Step 2: Mention Identification

- $p(e)$: MRF on CUIs

L sided hearing loss

HTN

Hyperlipidemia

Neurological deficits

...
Step 2: Mention Identification

- Problem: CUIs are latent variables on MIMIC (unlabeled)

- Variational learning, following:
  - Autoencoding Variational Bayes, Kingma and Welling, ICLR 2014
Step 2: Mention Identification

- **Objective:**
  - Maximize $\log(\sum_e p(m, e; \theta))$

- **Jensen’s inequality:**
  - $\forall q, \log(\sum_e p(m, e; \theta)) \geq \sum_e q(e|m, \xi) \log(\frac{p(m|e, \theta)}{q(e|m, \xi)})$

- **Joint maximization in $\xi, \theta$**
Step 2: Mention Identification

- Factorized $q$:
  - $q(e|m) = \prod_i q(e_i|m)$

Diagram: 

- L sided hearing loss
- HTN
- Hyper-lipidemia
- Neurological deficits
- ...

List: 

- L sided hearing loss
- HTN
- Hyper-lipidemia
- Neurological deficits
- ...
Step 2: Mention Identification

- Considers mention and neighbors:
  \[ q(e_i|m) = q(e_i|m_{i-2}, m_{i-1}, m_i, m_{i+1}, m_{i+2}) \]
Step 2: Mention Identification

- Neural network parameterization
- Semi-automated restricted support
- Supervised training gives 2^{nd} best accuracy on 2014 task
Step 2: Mention Identification

● Next steps:
  ● Pre-train parameters
  ● Use correlation model
  ● Train with variational algorithm
Review of Semi-Supervised methods

- Automatic labeling of data
- Label pre-selection
- Use prototypes
- Use features learned on larger corpus
Review of Semi-Supervised methods

- Domain adaptation: PubMed
- Multi-view learning
Review of Semi-Supervised methods

- Multi-view learning:
  - Other information on the patient: diagnosis codes, procedures, demographics, etc…
  - Jointly learn to predict those
Questions?