NLP and ML: Synergy or Divergence?

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay and IIT Patna

www.cse.iitb.ac.in/~pb

(21st Sept, 2016)
Roadmap

• Perspective
• Power of Data
• Some “lower level” NLP tasks
• Alignment in MT
• Annotation
• Cooperative WSD
• Sarcasm
• Conclusions
Perspective
NLP: a useful view

Increased Complexity Of Processing

Discourse and Coreference

Semantics

Parsing

Chunking

POS tagging

Morphology

Problem

Part of Speech Tagging

Morph Analysis

HMM

CRF

MEMM

Language

Hindi

Marathi

English

French

NLP Trinity

Algorithm

Semantics

Parsing
Why is NLP hard?
Ambiguity

- Lexical Ambiguity
- Structural Ambiguity
- Semantic Ambiguity
- Pragmatic Ambiguity
Examples

1. (ellipsis) Amsterdam airport: “Baby Changing Room”

2. (Attachment/grouping) Public demand changes (credit for the phrase: Jayant Haritsa):
   (a) Public demand changes, but does any body listen to them?
   (b) Public demand changes, and we companies have to adapt to such changes.
   (c) Public demand changes have pushed many companies out of business

3. (Attachment) Ishant ruled out of first test with Chickengunia (ToI: 21/9/16)

3. (Pragmatics-1) The use of shin bone is to locate furniture in a dark room

4. (Pragmatics-2) Blood flows on streets of Dhaka on Eid after animal sacrifice
New words and terms (people are very creative!!)

1. **ROFL**: rolling on the floor laughing; **LOL**: laugh out loud

2. **facebook**: to use facebook; **google**: to search

3. **communifake**: faking to talk on mobile; **Obamacare**: medical care system introduced through the mediation of President Obama (portmanteau words)

Example: Humour

1. (for a student of mine)
   Student: my thesis is on unsupervised WSD
   Prof. Sivakumar: But I thought Pushpakan is supervising your thesis!

2. (ToI, 11/4/15)
   If money does not grow on trees, why do banks have branches?

3. (ToI 2/3/15)
   Q: Have you heard of the kidnapping in the school?
   A: no, he got up
NLP: compulsory Inter layer interaction (1/2)

Text-1: “I saw the boy with a telescope which he dropped accidentally”
Text-2: “I saw the boy with a telescope which I dropped accidentally

Text-1:
(S
  (NP (PRP I))
  (VP
    (VBD saw)
    (NP (DT the) (NN boy))
    (PP (IN with) (NP (NP (DT a) (NN telescope))
      (SBAR (WHNP (WDT which)) (S (NP (PRP I))
        (VP (VBD dropped)
          (ADVP (RB accidentally)))))))))
)

Text-2:
(S
  (NP (PRP I))
  (VP
    (VBD saw)
    (NP (DT the) (NN boy))
    (PP (IN with) (NP (NP (DT a) (NN telescope))
      (SBAR (WHNP (WDT which)) (S (NP (PRP he))
        (VP (VBD dropped) (ADVP (RB accidentally)))))))
))
Inter layer interaction (2/2)

Text-1: “I saw the boy with a telescope which he dropped accidentally”
Text-2: “I saw the boy with a telescope which I dropped accidentally

nsubj(saw-2, I-1)
root(ROOT-0, saw-2)
det(boy-4, the-3)
dobj(saw-2, boy-4)
det(telescope-7, a-6)
prep_with(saw-2, telescope-7)
dobj(dropped-10, telescope-7)
nsubj(dropped-10, I-9)
rcmod(telescope-7, dropped-10)
advmod(dropped-10, accidentally-11)

nsubj(saw-2, I-1)
root(ROOT-0, saw-2)
det(boy-4, the-3)
dobj(saw-2, boy-4)
det(telescope-7, a-6)
prep_with(saw-2, telescope-7)
dobj(dropped-10, telescope-7)
nsubj(dropped-10, he-9)
rcmod(telescope-7, dropped-10)
advmod(dropped-10, accidentally-11)
NLP: deal with multilinguality

Language Typology
Languages differ in expressing thoughts: Agglutination

- Finnish: “istahtaisinkohan”
- English: "I wonder if I should sit down for a while"

Analysis:
- ist + "sit", verb stem
- ahta + verb derivation morpheme, "to do something for a while"
- isi + conditional affix
- n + 1st person singular suffix
- ko + question particle
- han a particle for things like reminder (with declaratives) or "softening" (with questions and imperatives)
Some scientists say our mind+in happening thoughts dream+become see

Some scientists opine that whatever we see in dreams are thoughts encased in our unconscious mind.

So far we have meaningful units of text.

But, we needs lot of data to achieve good vocabulary coverage and probability estimates
Use character as basic unit

कुछ शास्त्र में ने कहा हमारे मन मस्सों वाली चिंता स्वप्न माना जाता है।

That's looks like a good start, given we have no linguistic knowledge

Though, we essentially threw away the notion of a word!

The basic units don’t convey any meaning!

Can we do better?
Let’s try something better

First segment the character stream into *akshar*

*ie.* Consonant-vowel+ combinations

वैज्ञानिकों → वै ज्ञा नि कों

Why?

- Character vocabulary very small, ambiguous translations
- Syllable as a basic unit of speech

Translation output

कुछ वैज्ञानिकों का कहना है कि हमारे मन में होने वाले चिंताओं स्वप्न से देख लेते हैं।

*We get even better results!*  

*But, these basic units aren’t meaningful either!!*
This works for many language pairs

(Kunchukuttan & Bhattacharyya, 2016)

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Word</th>
<th>Morph</th>
<th>Character</th>
<th>Orth-Syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td>bn</td>
<td>hi</td>
<td>31.23</td>
<td>32.17</td>
<td>27.95</td>
<td>33.46</td>
</tr>
<tr>
<td>kK</td>
<td>mr</td>
<td>21.39</td>
<td>22.81</td>
<td>19.83</td>
<td>23.53</td>
</tr>
<tr>
<td>ml</td>
<td>ta</td>
<td>6.52</td>
<td>7.61</td>
<td>4.50</td>
<td>7.86</td>
</tr>
<tr>
<td>hi</td>
<td>ml</td>
<td>8.49</td>
<td>9.23</td>
<td>6.28</td>
<td>10.45</td>
</tr>
<tr>
<td>ml</td>
<td>hi</td>
<td>15.23</td>
<td>17.08</td>
<td>12.33</td>
<td>18.50</td>
</tr>
<tr>
<td>pa</td>
<td>hi</td>
<td>68.96</td>
<td>71.29</td>
<td>71.26</td>
<td>72.51</td>
</tr>
<tr>
<td>te</td>
<td>ml</td>
<td>6.62</td>
<td>7.86</td>
<td>6.00</td>
<td>8.51</td>
</tr>
</tbody>
</table>

So, what’s happening?

Anoop Kunchukuttan, Pushpak Bhattacharyya. Orthographic Syllable as basic unit for SMT between Related Languages. EMNLP. 2016.
These language pairs exhibit the following properties

**Lexical Similarity**: Cognates, loan-words, lateral borrowings

**Structural Correspondence**: Similar word order and parse structures

**Morphological Isomorphism**: Correspondence between suffixes/post-positions in language pairs
Implicit use of linguistic knowledge

• This technique worked because the properties of lexical similarity, structural correspondence and morphological isomorphism hold between related languages.
• A linguistic understanding is needed to understand the applicability and viability of NLP techniques.
• Many SMT techniques which claim language independence use implicit linguistic knowledge (Bender, 2011).
  – Classical methods of POS tagging and n-gram modelling assume simple morphology and rigid word-order.

Two approaches to NLP: Knowledge Based and ML based

- Classical NLP
  - Linguist
  - Annotation driven
  - Text data
  - Corpus
  - Rules
- Statistical NLP
  - Computer
  - Rules/probabilities
Rules: when and when not

• When the phenomenon is understood AND expressed, rules are the way to go
• “Do not learn when you know!!”

• When the phenomenon “seems arbitrary” at the current state of knowledge, DATA is the only handle!

• Rely on machine learning to tease truth out of data

• Expectation not always met with 😞
Why is probability important for NLP

Choose amongst competing options
Impact of probability: Language modeling

Probabilities computed in the context of corpora

1. $P(\text{"The sun rises in the east"})$
2. $P(\text{"The sun rise in the east"})$
   • Less probable because of grammatical mistake.
3. $P(\text{The svn rises in the east})$
   • Less probable because of lexical mistake.
4. $P(\text{The sun rises in the west})$
   • Less probable because of semantic mistake.
Empiricism vs. Rationalism

- Ken Church, “A Pendulum Swung too Far”, LILT, 2011

  - Availability of huge amount of data: what to do with it?
  - 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
  - 1970s: Rationalism (Chomsky, Minsky)
  - 1990s: Empiricism (IBM Speech Group, AT & T)
  - 2010s: Return of Rationalism?

*Resource generation will play a vital role in this revival of rationalism*
Power of Data
Automatic image labeling
(Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, 2014)

Automatically captioned: “Two pizzas sitting on top of a stove top oven”
Automatic image labeling (cntd)

<table>
<thead>
<tr>
<th>Describes without errors</th>
<th>Describes with minor errors</th>
<th>Somewhat related to the image</th>
<th>Unrelated to the image</th>
</tr>
</thead>
<tbody>
<tr>
<td>A person riding a motorcycle on a dirt road.</td>
<td>Two dogs play in the grass.</td>
<td>A skateboarder does a trick on a ramp.</td>
<td>A dog is jumping to catch a frisbee.</td>
</tr>
<tr>
<td>A group of young people playing a game of frisbee.</td>
<td>Two hockey players are fighting over the puck.</td>
<td>A little girl in a pink hat is blowing bubbles.</td>
<td>A refrigerator filled with lots of food and drinks.</td>
</tr>
<tr>
<td>A herd of elephants walking across a dry grass field.</td>
<td>A close up of a cat laying on a couch.</td>
<td>A red motorcycle parked on the side of the road.</td>
<td>A yellow school bus parked in a parking lot.</td>
</tr>
</tbody>
</table>

21 Sept 2016  nlp-ml: fuss  27
Thought Reader!

“I am hungry now”
Main methodology

- Object A: extract parts and features
- Object B which is in correspondence with A: extract parts and features
- LEARN mappings of these features and parts
- Use in NEW situations: called DECODING
Some foundational NLP tasks
Part of Speech Tagging

• POS Tagging: attaches to each word in a sentence a part of speech tag from a given set of tags called the Tag-Set
• Standard Tag-set : Penn Treebank (for English).
Example

“_“ The_DT mechanisms_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP ._.

21 Sept 2016 nlp-ml: fuss 32
Where does POS tagging fit in

Increased Complexity Of Processing

- Discourse and Coreference
- Semantics Extraction
- Parsing
- Chunking
- POS tagging
- Morphology

21 Sept 2016
# Penn tag set

<table>
<thead>
<tr>
<th>CC</th>
<th>Coord Conjuncn</th>
<th>and, but, or</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>Cardinal number</td>
<td>one, two</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
<td>the, some</td>
</tr>
<tr>
<td>EX</td>
<td>Existential there</td>
<td>there</td>
</tr>
<tr>
<td>FW</td>
<td>Foreign Word</td>
<td>mon dieu</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition</td>
<td>of, in, by</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
<td>big</td>
</tr>
<tr>
<td>JJR</td>
<td>Adj., comparative</td>
<td>bigger</td>
</tr>
<tr>
<td>JJS</td>
<td>Adj., superlative</td>
<td>biggest</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
<td>1, One</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
<td>can, should</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, sing. or mass</td>
<td>dog</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
<td>dogs</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, sing.</td>
<td>Edinburgh</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>Orkneys</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
<td>all, both</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
<td>'s</td>
</tr>
<tr>
<td>PP</td>
<td>Personal pronoun</td>
<td>I, you, she</td>
</tr>
<tr>
<td>PP$</td>
<td>Possessive pronoun</td>
<td>my, one's</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
<td>quickly</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
<td>faster</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
<td>fastest</td>
</tr>
</tbody>
</table>
### Penn Tagset cntd.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>VB</td>
<td>Verb, base form subsumes imperatives, infinitives and subjunctives</td>
<td><em>I want to dance</em></td>
</tr>
<tr>
<td>VBD</td>
<td>Verb, past tense includes the conditional form of the verb to be</td>
<td><em>I went to dance</em></td>
</tr>
<tr>
<td>VBG</td>
<td>Verb, gerund or present participle</td>
<td><em>I went to dance parties</em></td>
</tr>
<tr>
<td>VBN</td>
<td>Verb, past participle</td>
<td></td>
</tr>
<tr>
<td>VBP</td>
<td>Verb, non-3rd person singular present</td>
<td></td>
</tr>
<tr>
<td>VBZ</td>
<td>Verb, 3rd person singular present</td>
<td></td>
</tr>
<tr>
<td>TO</td>
<td>to</td>
<td></td>
</tr>
</tbody>
</table>

#### Language Phenomena

1. Most English nouns can act as verbs
2. Noun plurals have the same form as 3p1n verbs

---

Indian Language Tag set: Noun

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Category</th>
<th>Label</th>
<th>Annotation Convention</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subtype (level 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subtype (level 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Noun</td>
<td>N</td>
<td>N</td>
<td>ladakaa, raajaa, kitaaba</td>
</tr>
<tr>
<td>1.1</td>
<td>Common</td>
<td>NN</td>
<td>N__NN</td>
<td>kitaaba, kalama, cashmaa</td>
</tr>
<tr>
<td>1.2</td>
<td>Proper</td>
<td>NNP</td>
<td>N__NNP</td>
<td>Mohan, ravi, rashmi</td>
</tr>
<tr>
<td>1.4</td>
<td>Nloc</td>
<td>NST</td>
<td>N__NST</td>
<td>Uupara, niice, aage,</td>
</tr>
</tbody>
</table>
**Argmax computation (1/2)**

Best tag sequence

= $T^*$

= $\arg\max P(T|W)$

= $\arg\max P(T)P(W|T)$ (by Baye’s Theorem)

$P(T) = P(t_0=t_1t_2 \ldots t_{n+1}=.)$

= $P(t_0)P(t_1|t_0)P(t_2|t_1t_0)P(t_3|t_2t_1t_0) \ldots$

$P(t_n|t_{n-1}t_{n-2} \ldots t_0)P(t_{n+1}|t_n t_{n-1} \ldots t_0)$

= $P(t_0)P(t_1|t_0)P(t_2|t_1) \ldots P(t_n|t_{n-1})P(t_{n+1}|t_n)$

= $\prod_{i=0}^{N+1} P(t_i|t_{i-1})$  Bigram Assumption
Argmax computation (2/2)

\[
P(W|T) = P(w_0|t_0-t_{n+1})P(w_1|w_0t_0-t_{n+1})P(w_2|w_1w_0t_0-t_{n+1}) \ldots
\]
\[
P(w_n|w_{0-n}t_0-t_{n+1})P(w_{n+1}|w_0w_nt_0-t_{n+1})
\]

Assumption: A word is determined completely by its tag. This is inspired by speech recognition

\[
= P(w_0|t_0)P(w_1|t_1) \ldots P(w_{n+1}|t_{n+1})
\]

\[
= P(w_i|t_i)
\]

\[
= \prod_{i=1}^{n+1} P(w_i|t_i) \quad \text{(Lexical Probability Assumption)}
\]
Machine Translation and Machine Learning
Why is MT difficult: Language Divergence

- Languages have different ways of expressing meaning
  - Lexico-Semantic Divergence
  - Structural Divergence

Our work on English-IL Language Divergence with illustrations from Hindi
(Dave, Parikh, Bhattacharyya, Journal of MT, 2002)
Kinds of MT Systems
(point of entry from source to the target text)
Simplified Vauquois

Interlingua Based Translation

Transfer Based Translation

Direct Translation
Taxonomy of MT systems

MT Approaches

Knowledge Based; Rule Based MT
- Interlingua Based
- Transfer Based

Data driven; Machine Learning Based
- Example Based MT (EBMT)
- Statistical MT

21 Sept 2016  nlp-ml: fuss  44
RBMT-EBMT-SMT spectrum: knowledge (rules) intensive to data (learning) intensive
Can and should choose level of transfer

- राजा को नमन करो (Hindi; Indo Aryan)
  raajaa ko naman karo
  HG: king to obeisance do
  Give obeisance to the king (English; Indo-Aryan)

- राजाला नमन करा (Marathi; Indo Aryan)
  raajaalaa naman karaa
  king_to obeisance do

- அரசரர் வணங்கு (Tamil; Dravidian)
  aracarai vanaNku
  king_to obeisance_do

- နေပင်းဗောင်ရှိ (Manipuri; Tibeto Burman)
  niNgthoubu
  khoirammu
  king_to obeisance do
transfer amongst different language families

<table>
<thead>
<tr>
<th>Language</th>
<th>Inflected Verb/Inflected verb complex</th>
<th>Inflected Noun/Inflected Noun chunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>give obeisance</td>
<td>To the king</td>
</tr>
<tr>
<td>Hindi</td>
<td>naman karo</td>
<td>raajaa ko</td>
</tr>
<tr>
<td>Marathi</td>
<td>naman karaa</td>
<td>raajaalaa</td>
</tr>
<tr>
<td>Tamil</td>
<td>vanaNku</td>
<td>aracarai</td>
</tr>
<tr>
<td>Manipuri</td>
<td>Khoirammu</td>
<td>niNgthoubu</td>
</tr>
</tbody>
</table>
Data driven translation: Czeck-English data

- [nesu]  “I carry”
- [ponese]  “He will carry”
- [nese]  “He carries”
- [nesou]  “They carry”
- [yedu]  “I drive”
- [plavou]  “They swim”
To translate …

• I will carry.
• They drive.
• He swims.
• They will drive.
<table>
<thead>
<tr>
<th>Hindi-English data</th>
</tr>
</thead>
<tbody>
<tr>
<td>[DhotA huM]</td>
</tr>
<tr>
<td>[DhoegA]</td>
</tr>
<tr>
<td>[DhotA hAi]</td>
</tr>
<tr>
<td>[Dhote hAi]</td>
</tr>
<tr>
<td>[chalAtA huM]</td>
</tr>
<tr>
<td>[tErte hEM]</td>
</tr>
</tbody>
</table>
Bangla-English data

- [bai] “I carry”
- [baibe] “He will carry”
- [bay] “He carries”
- [bay] “They carry”
- [chAlAi] “I drive”
- [sAMtrAy] “They swim”
Word alignment as the crux of Statistical Machine Translation

**English**

(1) three rabbits  
(a) b

(2) rabbits of Grenoble  
b (c) d

**French**

(1) trois lapins  
(w) x

(2) lapins de Grenoble  
x (y) z

21 Sept 2016  
nlp-ml: fuss  
52
Initial Probabilities:
each cell denotes $t(a \leftrightarrow w), \ t(a \leftrightarrow x)$ etc.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>w</strong></td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td><strong>x</strong></td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td><strong>y</strong></td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td><strong>z</strong></td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
</tbody>
</table>
### “counts”

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### counts

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Revised probabilities table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/2</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1/2</td>
<td>5/12</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>1/6</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>1/6</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>
## "revised counts"

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/2</td>
<td>3/8</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1/2</td>
<td>5/8</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>0</td>
<td>5/9</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td>0</td>
<td>2/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

21 Sept 2016 nlp-ml. fuss 56
Re-Revised probabilities table

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1/2</td>
<td>3/16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>1/2</td>
<td>85/144</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>1/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>1/9</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Continue until convergence; notice that (b, x) binding gets progressively stronger; b=rabbits, x=lapins
Derivation: Key Notations

English vocabulary : $V_E$
French vocabulary : $V_F$
No. of observations / sentence pairs : $S$
Data $D$ which consists of $S$ observations looks like,

\[ e^1_1, e^1_2, \ldots, e^1_l \Leftrightarrow f^1_1, f^1_2, \ldots, f^1_m \]
\[ e^2_1, e^2_2, \ldots, e^2_l \Leftrightarrow f^2_1, f^2_2, \ldots, f^2_m \]
\[ \ldots \]
\[ e^s_1, e^s_2, \ldots, e^s_l \Leftrightarrow f^s_1, f^s_2, \ldots, f^s_m \]
\[ \ldots \]
\[ e^s_1, e^s_2, \ldots, e^s_l \Leftrightarrow f^s_1, f^s_2, \ldots, f^s_m \]

No. words on English side in $s^{th}$ sentence : $l^s$
No. words on French side in $s^{th}$ sentence : $m^s$

\[ index_E(e^s_p) = \text{Index of English word } e^s_p \text{ in English vocabulary/dictionary} \]
\[ index_F(f^s_q) = \text{Index of French word } f^s_q \text{ in French vocabulary/dictionary} \]

(Thanks to Sachin Pawar for helping with the maths formulae processing)

21 Sept 2016    nlp-ml: fuss    58
Modeling: Hidden variables and parameters

Hidden Variables (Z):
Total no. of hidden variables = $\sum_{s=1}^{S} l^s m^s$ where each hidden variable is as follows:
$z_{pq}^s = 1$, if in $s^{th}$ sentence, $p^{th}$ English word is mapped to $q^{th}$ French word.
$z_{pq}^s = 0$, otherwise

Parameters (Θ):
Total no. of parameters = $|V_E| \times |V_F|$, where each parameter is as follows:
$P_{i,j} = $ Probability that $i^{th}$ word in English vocabulary is mapped to $j^{th}$ word in French vocabulary
Likelihoods

Data Likelihood $L(D; \Theta)$:

$$L(D; \Theta) = \prod_{s=1}^{S} \prod_{p=1}^{l^s} \prod_{q=1}^{m^s} \left( P_{\text{index}_E(e^s_p), \text{index}_F(f^s_q)} \right)^{z^s_{pq}}$$

Data Log-Likelihood $LL(D; \Theta)$:

$$LL(D; \Theta) = \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} z^s_{pq} \log \left( P_{\text{index}_E(e^s_p), \text{index}_F(f^s_q)} \right)$$

Expected value of Data Log-Likelihood $E(LL(D; \Theta))$:

$$E(LL(D; \Theta)) = \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} E(z^s_{pq}) \log \left( P_{\text{index}_E(e^s_p), \text{index}_F(f^s_q)} \right)$$
Constraint and Lagrangian

\[ \sum_{j=1}^{|V_F|} P_{i,j} = 1 \quad \forall i \]

\[ \sum_{s=1}^{S} \sum_{p=1}^{l_s} \sum_{q=1}^{m_s} E(z_{pq}^s) \log \left( P_{index_B(e_p^s),index_F(f_q^s)} \right) - \sum_{i=1}^{|V_E|} \lambda_i \left( \sum_{j=1}^{|V_F|} P_{i,j} - 1 \right) \]
Differentiating wrt $P_{ij}$

$$
\sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{\text{index}_E(e_p^s),i} \delta_{\text{index}_F(f_q^s),j} \left( \frac{E(z_{pq}^s)}{P_{i,j}} \right) - \lambda_i = 0
$$

$$
P_{i,j} = \frac{1}{\lambda_i} \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{\text{index}_E(e_p^s),i} \delta_{\text{index}_F(f_q^s),j} E(z_{pq}^s)
$$

$$
\sum_{j=1}^{|V_F|} P_{i,j} = 1 = \sum_{j=1}^{|V_F|} \frac{1}{\lambda_i} \sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} \delta_{\text{index}_E(e_p^s),i} \delta_{\text{index}_F(f_q^s),j} E(z_{pq}^s)
$$

21 Sept 2016  nlp-ml: fuss  62
Final E and M steps

M-step

\[ P_{i,j} = \frac{\sum_{s=1}^{S} \sum_{p=1}^{l_s} \sum_{q=1}^{m_s} \delta_{\text{index}_E(e_p^s)} \cdot \delta_{\text{index}_F(f_q^s)} \cdot j \cdot E(z_{pq}^s)}{\sum_{j=1}^{|V_F|} \sum_{s=1}^{S} \sum_{p=1}^{l_s} \sum_{q=1}^{m_s} \delta_{\text{index}_E(e_p^s)} \cdot \delta_{\text{index}_F(f_q^s)} \cdot j \cdot E(z_{pq}^s)}, \forall i, j \]

E-step

\[ E(z_{pq}^s) = \frac{P_{\text{index}_E(e_p^s), \text{index}_F(f_q^s)}}{\sum_{q'=1}^{m_s} P_{\text{index}_E(e_p^s), \text{index}_F(f_{q'}^s)}}, \forall s, p, q \]
A recent study

PAN Indian SMT
(Anoop K, Abhijit Mishra, Pushpak Bhattacharyya, LREC 2014)
Natural Partitioning of SMT systems

<table>
<thead>
<tr>
<th></th>
<th>hi</th>
<th>ur</th>
<th>pa</th>
<th>bn</th>
<th>gu</th>
<th>mr</th>
<th>kK</th>
<th>ta</th>
<th>te</th>
<th>ml</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>hi</td>
<td>61.28</td>
<td>68.21</td>
<td>34.96</td>
<td>51.31</td>
<td>39.12</td>
<td>37.81</td>
<td>14.43</td>
<td>21.38</td>
<td>10.98</td>
<td>29.23</td>
<td></td>
</tr>
<tr>
<td>ur</td>
<td>61.42</td>
<td>52.02</td>
<td>29.59</td>
<td>39.00</td>
<td>27.57</td>
<td>28.29</td>
<td>11.95</td>
<td>16.61</td>
<td>8.65</td>
<td>22.46</td>
<td></td>
</tr>
<tr>
<td>pa</td>
<td>73.31</td>
<td>56.00</td>
<td>29.89</td>
<td>43.85</td>
<td>30.87</td>
<td>30.72</td>
<td>10.75</td>
<td>18.81</td>
<td>9.11</td>
<td>23.83</td>
<td></td>
</tr>
<tr>
<td>bn</td>
<td>37.69</td>
<td>32.08</td>
<td>31.38</td>
<td>28.14</td>
<td>22.09</td>
<td>23.47</td>
<td>10.94</td>
<td>13.40</td>
<td>8.10</td>
<td>18.76</td>
<td></td>
</tr>
<tr>
<td>gu</td>
<td>55.66</td>
<td>44.12</td>
<td>45.14</td>
<td>28.50</td>
<td>32.06</td>
<td>30.48</td>
<td>12.57</td>
<td>17.22</td>
<td>8.01</td>
<td>19.78</td>
<td></td>
</tr>
<tr>
<td>mr</td>
<td>45.11</td>
<td>32.60</td>
<td>33.28</td>
<td>23.73</td>
<td>32.42</td>
<td>27.81</td>
<td>10.74</td>
<td>12.89</td>
<td>7.65</td>
<td>17.62</td>
<td></td>
</tr>
<tr>
<td>kK</td>
<td>41.92</td>
<td>34.00</td>
<td>34.31</td>
<td>24.59</td>
<td>31.07</td>
<td>27.52</td>
<td>10.36</td>
<td>14.80</td>
<td>7.89</td>
<td>17.07</td>
<td></td>
</tr>
<tr>
<td>ta</td>
<td>20.48</td>
<td>18.12</td>
<td>15.57</td>
<td>13.21</td>
<td>16.53</td>
<td>11.60</td>
<td>11.87</td>
<td>8.48</td>
<td>6.31</td>
<td>11.79</td>
<td></td>
</tr>
<tr>
<td>te</td>
<td>28.88</td>
<td>25.07</td>
<td>25.56</td>
<td>16.57</td>
<td>20.96</td>
<td>14.94</td>
<td>17.27</td>
<td>8.68</td>
<td>6.68</td>
<td>12.34</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>28.94</td>
<td>22.96</td>
<td>22.33</td>
<td>15.33</td>
<td>15.44</td>
<td>12.11</td>
<td>13.66</td>
<td>6.43</td>
<td>6.55</td>
<td>4.65</td>
<td></td>
</tr>
</tbody>
</table>

Baseline PBSMT - % BLEU scores (S1)

- Clear partitioning of translation pairs by language family pairs, based on translation accuracy.
  - Shared characteristics within language families make translation simpler
  - Divergences among language families make translation difficult
Using Bridge to mitigate resource scarcity
L1 $\rightarrow$ bridge $\rightarrow$ L2 (Wu and Wang 2009)

- Resource rich and resource poor language pairs
- Question-1: How about translating through a ‘bridge’?
- Question-2: how to choose the bridge?
Mathematical preliminaries

\[ e_{\text{best}} = \arg \max_e p(e|f) \]
\[ = \arg \max_e p(f|e)p_{\text{LM}}(e) \]

Where \( p(f|e) \) is given by:

\[ p(f|e) = p\left( \frac{f}{e} \right) = \prod_{i=1}^{I} \phi \left( \frac{\bar{f}_i}{\bar{e}_i} \right) d(a_i - b_{i-1}) p_w \left( \frac{\bar{f}_i}{\bar{e}_i, a_i} \right) \]

\[ \phi \left( \frac{\bar{f}_i}{\bar{e}_i} \right) = \sum_{\bar{p}_i} \phi \left( \frac{\bar{f}_i}{\bar{p}_i} \right) \phi \left( \frac{\bar{p}_i}{\bar{e}_i} \right) \]

\[ p_w \left( \frac{\bar{f}_i}{\bar{a}_i} \right) = \prod_{l=1}^{n} \frac{1}{|m/(l,m) \in a|} \sum_{\forall (l,m) \in a} w(f_l/e_l) \]
Triangulation approach

- Important to induce language dependent components such as phrase translation probability and lexical weight
## English-Hindi SMT: Resource Details

<table>
<thead>
<tr>
<th>Segment</th>
<th>#Sentences</th>
<th>#Unique Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>46277</td>
<td>39452 (en), 41418 (hi)</td>
</tr>
<tr>
<td>Tuning</td>
<td>500</td>
<td>2623 (en), 2816 (hi)</td>
</tr>
<tr>
<td>Test</td>
<td>2000</td>
<td>6722 (en), 7102 (hi)</td>
</tr>
<tr>
<td>Monolingual</td>
<td>1538429</td>
<td>558206</td>
</tr>
</tbody>
</table>

21 Sept 2016  nlp-ml: fuss  69
<table>
<thead>
<tr>
<th>Model</th>
<th>l=1k</th>
<th>l=2k</th>
<th>l=3k</th>
<th>l=4k</th>
<th>l=5k</th>
<th>l=6k</th>
<th>l=7k</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT_l</td>
<td>8.86</td>
<td>11.39</td>
<td>13.78</td>
<td>15.62</td>
<td>16.78</td>
<td>18.03</td>
<td>19.02</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_BN</td>
<td>14.34</td>
<td>16.51</td>
<td>17.87</td>
<td>18.72</td>
<td>19.79</td>
<td>20.45</td>
<td>21.14</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_GU</td>
<td>13.91</td>
<td>16.15</td>
<td>17.38</td>
<td>18.77</td>
<td>19.65</td>
<td>20.46</td>
<td>21.17</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_KK</td>
<td>13.68</td>
<td>15.88</td>
<td>17.3</td>
<td>18.33</td>
<td>19.21</td>
<td>20.1</td>
<td>20.51</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_ML</td>
<td>11.22</td>
<td>13.04</td>
<td>14.71</td>
<td>15.91</td>
<td>17.02</td>
<td>17.76</td>
<td>18.72</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_MA</td>
<td>13.3</td>
<td>15.27</td>
<td>16.71</td>
<td>18.13</td>
<td>18.9</td>
<td>19.49</td>
<td>20.07</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_PU</td>
<td>15.63</td>
<td>17.62</td>
<td>18.77</td>
<td>19.88</td>
<td>20.76</td>
<td>21.53</td>
<td>22.01</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_TA</td>
<td>12.36</td>
<td>14.09</td>
<td>15.73</td>
<td>16.97</td>
<td>17.77</td>
<td>18.23</td>
<td>18.85</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_UR</td>
<td>15.34</td>
<td>17.37</td>
<td>18.36</td>
<td>19.35</td>
<td>20.46</td>
<td>21.14</td>
<td>21.35</td>
</tr>
<tr>
<td>DIRECT_l+BRIDGE_PU_UR</td>
<td>20.53</td>
<td>21.3</td>
<td>21.97</td>
<td>22.58</td>
<td>22.64</td>
<td>22.98</td>
<td>24.73</td>
</tr>
</tbody>
</table>

21 Sept 2016  nlp-ml: fuss  70
Effect of Multiple Pivots

Fr-Es translation using 2 pivots

![Graph showing BLEU scores for Fr-Es translation with different systems and pivot strategies.]

- Interpolated-En+De
- Interpolated-En
- Interpolated-De
- Standard
- Pivot-En+De

Fr-Es Data (k Pairs)

<table>
<thead>
<tr>
<th>System</th>
<th>Ja→Hi</th>
<th>Hi→Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>33.86</td>
<td>37.47</td>
</tr>
<tr>
<td>Direct+best pivot</td>
<td>35.74 (es)</td>
<td>39.49 (ko)</td>
</tr>
<tr>
<td>Direct+Best-3 pivots</td>
<td>38.22</td>
<td>41.09</td>
</tr>
<tr>
<td>Direct+All 7 pivots</td>
<td>38.42</td>
<td>40.09</td>
</tr>
</tbody>
</table>

Hi-Ja translation using 7 pivots

Annotation
Definition

(Edvard Hovy, ACL 2010, tutorial on annotation)

• Annotation (‘tagging’) is the process of adding new information into raw data by human annotators.

• Typical annotation steps:
  – Decide which fragment of the data to annotate
  – Add to that fragment a specific bit of information
  – chosen from a fixed set of options
(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

(According to a new research, those people who have a busy social life, have larger space in a part of their brain).
Ambiguity of लोगों (People)

- लोग, जन, लोक, जनमानस, पब्लिक - एक से अधिक व्यक्ति "लोगों के हित में काम करना चाहिए"
  - (English synset) multitude, masses, mass, hoi_polloi, people, the_great_unwashed - the common people generally "separate the warriors from the mass" "power to the people"

- दुनिया, दुनियाँ, संसार, विश्व, जगत, जहाँ, जहान, ज़माना, जमाना, लोक, दुनियावाले, दुनियाँवाले, लोग - संसार में रहने वाले लोग "महात्मा गांधी का सम्मान पूरी दुनिया करती है / मैं इस दुनिया की परवाह नहीं करता / आज की दुनिया पैसे के पीछे भाग रही है"
  - (English synset) populace, public, world - people in general considered as a whole "he is a hero in the eyes of the public"
"My dog also likes eating sausage."

(ROOT
 (S
  (NP
    (PRP$ My) (NN dog))
  (ADVP (RB also))
  (VP (VBZ likes)
    (S (VP (VBG eating)
      (NP (NN sausage)))))) (. .)))

poss(dog-2, My-1)
nsubj(likes-4, dog-2)
advmod(likes-4, also-3)
root(ROOT-0, likes-4)
xcomp(likes-4, eating-5)
dobj(eating-5, sausage-6)
Good annotators and good annotation designers are rare to find

- An annotator has to understand BOTH language phenomena and the data
- An annotation designer has to understand BOTH linguistics and statistics!
Scale of effort involved in annotation (1/2)

- Penn Treebank
  - Total effort: 8 million words, 20-25 man years (5 persons for 4-5 years)
- Ontonotes: Annotate 300K words per year (1 person per year)
  - news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows,
  - with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)
  - in English, Chinese, and Arabic
- Prague Discourse Treebank (Czech): 500,000 words, 20-25 man years (4-5 persons for 5 years)
Sense marked corpora created at IIT Bombay

- http://www.cfilt.iitb.ac.in/wsd/annotated_corpus
- English: Tourism (~170000), Health (~150000)
- Hindi: Tourism (~170000), Health (~80000)
- Marathi: Tourism (~120000), Health (~50000)
  - 6 man years for each <L,D> combination (3 persons for 2 years)
Serious world wide effort on leveraging multiliguality

- Greg Durrett, Adam Pauls, and Dan Klein, *Syntactic Transfer Using Bilingual Lexicon*, EMNLP-CoNLL, 2012
- Dipanjan Das and Slav Petrov, *Unsupervised Part of Speech Tagging with Bilingual Graph-Based Projections*, ACL, 2011
Cooperative Word Sense Disambiguation
Definition: WSD

- Given a context:
  - Get “meaning”s of
    - a set of words (targetted wsd)
    - or all words (all words wsd)
  - The “Meaning” is usually given by the id of senses in a sense repository
    - usually the wordnet
Example: “operation” (from Princeton Wordnet)

- **Operation**, surgery, surgical operation, surgical procedure, surgical process -- (a medical procedure involving an incision with instruments; performed to repair damage or arrest disease in a living body; "they will schedule the operation as soon as an operating room is available"; "he died while undergoing surgery")  TOPIC->(noun) surgery#1

- **Operation**, military operation -- (activity by a military or naval force (as a maneuver or campaign); "it was a joint operation of the navy and air force")  TOPIC->(noun) military#1, armed forces#1, armed services#1, military machine#1, war machine#1

- mathematical process, mathematical **operation**, operation -- ((mathematics) calculation by mathematical methods; "the problems at the end of the chapter demonstrated the mathematical processes involved in the derivation"; "they were learning the basic operations of arithmetic")  TOPIC->(noun) mathematics#1, math#1, maths#1
WSD for ALL Indian languages: Critical resource: INDOWORDNET
Synset Based Multilingual Dictionary

<table>
<thead>
<tr>
<th>Concepts</th>
<th>L1 (English)</th>
<th>L2 (Hindi)</th>
<th>L3 (Marathi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a youthful male person</td>
<td>{malechild, boy}</td>
<td>लड़का (ladkaa), वालक (bvalak), बच्चा (bachcha)</td>
<td>मुलगा (mulga), पारगा (parga), पोर (por)</td>
</tr>
</tbody>
</table>

A sample entry from the MultiDict

- Expansion approach for creating wordnets [Mohanty et. al., 2008]
- Instead of creating from scratch link to the synsets of existing wordnet
- Relations get borrowed from existing wordnet
Cross Linkages Between Synset Members

- Captures native speakers intuition
- Wherever the word *ladkaa* appears in Hindi one would expect to see the word *mulgaa* in Marathi
- A few wordnet pairs do not have explicit word linkages within synset, in which case one assumes every word is linked all words on the other side
Resources for WSD - wordnet and corpora: 5 scenarios

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Annotated Corpus in L1</th>
<th>Aligned Wordnets</th>
<th>Annotated Corpus in L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>✓</td>
<td>✓</td>
<td>Varies</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>❌</td>
<td>✓</td>
<td>❌</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>Seed</td>
<td>✓</td>
<td>Seed</td>
</tr>
</tbody>
</table>
Unsupervised WSD
(No annotated corpora)

Khapra, Joshi and Bhattacharyya, IJCNLP 2011
ESTIMATING SENSE DISTRIBUTIONS

If sense tagged Marathi corpus were available, we could have estimated

\[ P(S_1^{mar}|maan) = \frac{\#(S_1^{mar}, maan)}{\#(S_1^{mar}, maan) + \#(S_2^{mar}, maan)} \]

But such a corpus is not available
EM for estimating sense distributions

\[
P(S_{1}^{mar}|maan) = \frac{\#(gardan) + \#(gala)}{\#(gardan) + \#(gala) + \frac{\#(maan)}{P(S_{1}^{mar}|maan)} + P(S_{1}^{mar}|greeva) \cdot \#(greeva) + P(S_{3}^{mar}|awaaj) \cdot \#(awaaj) + P(S_{3}^{mar}|swar) \cdot \#(swar)}
\]

\[
P(S_{1}^{hin}|gala) = \frac{P(S_{1}^{mar}|gala) \cdot \#(gala) + P(S_{1}^{mar}|greeva) \cdot \#(greeva) + P(S_{3}^{mar}|awaaj) \cdot \#(awaaj) + P(S_{3}^{mar}|swar) \cdot \#(swar)}{\#(gardan) + \#(gala) + \frac{\#(maan)}{P(S_{1}^{mar}|maan)} + P(S_{1}^{mar}|greeva) \cdot \#(greeva) + P(S_{3}^{mar}|awaaj) \cdot \#(awaaj) + P(S_{3}^{mar}|swar) \cdot \#(swar)}
\]
## Results & Discussions

- Performance of projection using manual cross linkages is within 7% of Self-Training.
- Performance of projection using probabilistic cross linkages is within 10-12% of Self-Training – remarkable since no additional cost incurred in target language.
- Both MCL and PCL give 10-14% improvement over Wordnet First Sense Baseline.
- *Not prudent to stick to knowledge based and unsupervised approaches – they come nowhere close to MCL or PCL.*

### Table:

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Tourism</th>
<th>Health</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P%</td>
<td>R%</td>
</tr>
<tr>
<td>MCL</td>
<td>73.36</td>
<td>68.83</td>
</tr>
<tr>
<td>PCL</td>
<td>68.57</td>
<td>67.93</td>
</tr>
<tr>
<td>IWSD-Self</td>
<td>78.36</td>
<td>77.77</td>
</tr>
<tr>
<td>WFS</td>
<td>57.15</td>
<td>57.15</td>
</tr>
<tr>
<td>PPR</td>
<td>51.49</td>
<td>51.49</td>
</tr>
<tr>
<td>Unsup</td>
<td>9.01</td>
<td>9.01</td>
</tr>
</tbody>
</table>

Our values:

- Manual Cross Linkages
- Probabilistic Cross Linkages
- Skyline - self training data is available
- Wordnet first sense baseline
- S-O-T-A Knowledge Based Approach
- S-O-T-A Unsupervised Approach
Harnessing Context Incongruity for Sarcasm Detection

1. Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya, Harnessing Context Incongruity for Sarcasm Detection, ACL 2015

2. Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman, Are Word Embedding-based Features Useful for Sarcasm Detection?, EMNLP 2016
Goal

The relationship between context incongruity and sarcasm has been studied in linguistics.

We present a statistical system that harnesses context incongruity as a basis for sarcasm detection in the form of two kinds of incongruity features: explicit and implicit.
Incongruity is defined as ‘the state of being not in agreement, as with principles’.

Ivanko and Pexman (2003) state that the sarcasm processing time (time taken by humans to understand sarcasm) depends on the degree of context incongruity between the statement and the context.
Two kinds of incongruity

• Explicit incongruity
  – Overtly expressed through sentiment words of both polarities
  – Contribute to almost 11% of sarcasm instances
    ‘I love being ignored’

• Implicit incongruity
  – Covertly expressed through phrases of implied sentiment
    ‘I love this paper so much that I made a doggy bag out of it’
# Feature Set

## Lexical

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>Unigrams in the training corpus</td>
</tr>
</tbody>
</table>

## Pragmatic

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capitalization</td>
<td>Numeric feature indicating presence of capital letters</td>
</tr>
<tr>
<td>Emoticons &amp; laughter expressions</td>
<td>Numeric feature indicating presence of emoticons and ‘lol’s</td>
</tr>
<tr>
<td>Punctuation marks</td>
<td>Numeric feature indicating presence of punctuation marks</td>
</tr>
</tbody>
</table>

**Implicit Incongruity** *(Based on Riloff et al)*

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit Sentiment Phrases</td>
<td>Boolean feature indicating phrases extracted from the implicit phrase extraction step</td>
</tr>
</tbody>
</table>

**Explicit Incongruity** *(Based on Ramteke et al)*

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Explicit incongruity</td>
<td>Number of times a word is followed by a word of opposite polarity</td>
</tr>
<tr>
<td>Largest positive /negative subsequence</td>
<td>Length of largest series of words with polarity unchanged</td>
</tr>
<tr>
<td>#Positive words</td>
<td>Number of positive words</td>
</tr>
<tr>
<td>#Negative words</td>
<td>Number of negative words</td>
</tr>
<tr>
<td>Lexical Polarity</td>
<td>Polarity of a tweet based on words present</td>
</tr>
</tbody>
</table>
## Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Text-form</th>
<th>Method of labeling</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet-A</td>
<td>Tweets</td>
<td>Using sarcasm-based hashtags as labels</td>
<td>5208 total, 4170 sarcastic</td>
</tr>
<tr>
<td>Tweet-B</td>
<td>Tweets</td>
<td>Manually labeled (Given by Riloff et al (2013))</td>
<td>2278 total, 506 sarcastic</td>
</tr>
<tr>
<td>Discussion-A</td>
<td>Discussion forum posts (IAC Corpus)</td>
<td>Manually labeled (Given by Walker et al (2012))</td>
<td>1502 total, 752 sarcastic</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Features</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Algorithm by Riloff et al. (2013)</td>
<td>0.774</td>
<td>0.098</td>
<td>0.173</td>
</tr>
<tr>
<td>Ordered</td>
<td>0.799</td>
<td>0.337</td>
<td>0.474</td>
</tr>
<tr>
<td>Unordered</td>
<td>0.799</td>
<td>0.337</td>
<td>0.474</td>
</tr>
<tr>
<td><strong>Our system</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical (Baseline)</td>
<td>0.820</td>
<td>0.867</td>
<td>0.842</td>
</tr>
<tr>
<td>Lexical+Implicit</td>
<td>0.822</td>
<td>0.887</td>
<td>0.853</td>
</tr>
<tr>
<td>Lexical+Explicit</td>
<td>0.807</td>
<td>0.985</td>
<td>0.8871</td>
</tr>
<tr>
<td>All features</td>
<td>0.814</td>
<td>0.976</td>
<td>0.8876</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approach</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riloff et al. (2013) (best reported)</td>
<td>0.62</td>
<td>0.44</td>
<td>0.51</td>
</tr>
<tr>
<td>Maynard and Greenwood (2014)</td>
<td>0.46</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>Our system (all features)</td>
<td>0.77</td>
<td>0.51</td>
<td>0.61</td>
</tr>
</tbody>
</table>

**Tweet-A**

**Tweet-B**

**Discussion-A**

<table>
<thead>
<tr>
<th>Features</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical (Baseline)</td>
<td>0.645</td>
<td>0.508</td>
<td>0.568</td>
</tr>
<tr>
<td>Lexical+Explicit</td>
<td>0.698</td>
<td>0.391</td>
<td>0.488</td>
</tr>
<tr>
<td>Lexical+Implicit</td>
<td>0.513</td>
<td>0.762</td>
<td>0.581</td>
</tr>
<tr>
<td>All features</td>
<td>0.489</td>
<td>0.924</td>
<td>0.640</td>
</tr>
</tbody>
</table>
When explicit incongruity is absent

A woman needs a man like a fish needs bicycle

Word2Vec similarity(man, woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

Can word embedding-based features when augmented to features reported in prior work improve the performance of sarcasm detection?
Word embedding-based features

(Stop words removed)

**Unweighted similarity features (S):**
For every word and word pair,
1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

**Distance-weighted similarity features (WS):** 4 S features weighted by linear distance between the two words

<table>
<thead>
<tr>
<th></th>
<th>man</th>
<th>woman</th>
<th>fish</th>
<th>needs</th>
<th>bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>-</td>
<td>0.766</td>
<td>0.151</td>
<td>0.078</td>
<td>0.229</td>
</tr>
<tr>
<td>woman</td>
<td>0.766</td>
<td>-</td>
<td>0.084</td>
<td>0.060</td>
<td>0.229</td>
</tr>
<tr>
<td>fish</td>
<td>0.151</td>
<td>0.084</td>
<td>-</td>
<td>0.022</td>
<td>0.130</td>
</tr>
<tr>
<td>needs</td>
<td>0.078</td>
<td>0.060</td>
<td>0.022</td>
<td>-</td>
<td>0.060</td>
</tr>
<tr>
<td>bicycle</td>
<td>0.229</td>
<td>0.229</td>
<td>0.130</td>
<td>0.060</td>
<td>-</td>
</tr>
</tbody>
</table>

**Both (S+WS):** 8 features
Experiment setup

- **Dataset**: 3629 Book snippets (759 sarcastic) downloaded from GoodReads website. Labeled by users with tags. We download ones with ‘sarcasm’ as sarcastic, ones with ‘philosophy’ as non-sarcastic.

- Five-fold cross-validation

- **Classifier**: SVM-Perf optimised for F-score

- **Configurations**:
  - Four prior works (augmented with our sets of features)
  - Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)
Results (1/2)

<table>
<thead>
<tr>
<th>Features</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unigrams</td>
<td>67.2</td>
<td>78.8</td>
<td>72.53</td>
</tr>
<tr>
<td>S</td>
<td>64.6</td>
<td>75.2</td>
<td>69.49</td>
</tr>
<tr>
<td>WS</td>
<td>67.6</td>
<td>51.2</td>
<td>58.26</td>
</tr>
<tr>
<td>Both</td>
<td>67</td>
<td>52.8</td>
<td>59.05</td>
</tr>
</tbody>
</table>

- **Observation**: Only word embedding-based features will not suffice. ‘Augmentation’ to other known useful features necessary.
- **Observation**: Using word embedding-based features improves sarcasm detection, for multiple word embedding types and feature sets.
Multiword Expressions

About half the lexical items in most languages are multiwords!
Multi-Word Expressions (MWE)

– Necessary Condition
  • Word sequence separated by space/delimiter

– Sufficient Conditions
  • Non-compositionality of meaning
  • Fixity of expression
    – In lexical items
    – In structure and order
Examples – Necessary condition

• Non-MWE example:
  – Marathi: सरकार हक्काबक्का झाले
  – Roman: sarakAra HakkAbakkA JZAle
  – Meaning: government was surprised

• MWE example:
  – Hindi: गरीब नवाज़
  – Roman: garlba navAjZa
  – Meaning: who nourishes poor
Examples - Sufficient conditions (Non-compositionality of meaning)

- Konkani: पोटांत चाबता
  Roman: poTAMta cAbatA
  Meaning: to feel jealous

- Telugu: చెట్టు కిందిక ప్లీడరు
  Roman: ceVttu kiMXa pLIdaru
  Meaning: an idle person

- Bangla: মাটির মানুষ
  Roman: mAtira mAnuSa
  Meaning: a simple person/son of the soil
Examples – Sufficient conditions (Fixity of expression)

In lexical items

• Hindi
  – usane muJe gAIl dl
  – *usane muJe gAIl pradAna kl
• Bangla
  – jabajjIbana karadaMda
  – *jIbanabhara karadaMda
  – *jabajjIbana jela

• English (1)
  – life imprisonment
  – *lifelong imprisonment
• English (2)
  – Many thanks
  – *Plenty of thanks
Examples – Sufficient conditions
(In structure and order)

- **English example**
  - kicked the bucket (died)
  - the bucket was kicked
    (not passivizable in the sense of dying)

- **Hindi example**
  - उम्र कैद
  - umra kEda (life imprisonment)
  - umra bhara kEda
### MW task (NLP + ML)

<table>
<thead>
<tr>
<th>String + Morph</th>
<th>POS</th>
<th>POS + WN</th>
<th>POS + List</th>
<th>Chunking</th>
<th>Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rules</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Onomaetopic Reduplication (tik tik, chham chham)</td>
<td>Non-</td>
<td>Non-re</td>
<td>Conjunct verb (verbalizer list), Compound verb (vector verb list)</td>
<td>Non-contiguous something</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Onomaetopic Reduplication (ghar ghar)</td>
<td>dup (Syn, Anto, Hypo) (raat din, dhan doulat)</td>
<td>(salaha dena, has uthama)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Statistical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collections or fixed expressions (many thanks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Idioms will be list morph + look up**
Summary

• POS tagging: done by ML predominantly

• Alignment in MT: predominantly ML; but cannot do without linguistics when facing rich morphology

• Co-operative WSD
  – Good linguistics (high quality linked wordnets) + Good ML (novel EM formulation)

• Sarcasm (difficult sentiment analysis problem)
  – Good NLP (incongruity) + good ML (string kernels?)

• MWE processing: FIXITY or colocation: ML is the only way; no apparent reason for fixity.
Conclusions

• Both Linguistics and Computation needed: Linguistics is the eye, Computation the body

• It is possible to leverage the resources created for one language in another

• Language phenomenon → Formalization → Hypothesis formation → Experimentation → Interpretation (Natural Science like flavor)

• Theory=Linguistics+NLP, Technique=ML
URLS

(publications) http://www.cse.iitb.ac.in/~pb

(resources) http://www.cfilt.iitb.ac.in
Thank you

Questions?
Word embedding-based features for sarcasm detection
(To appear in EMNLP 2016)

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman,
Are Word Embedding-based Features Useful for Sarcasm Detection?, EMNLP 2016, Austin, Texas, USA, November 1-5, 2016.
Introduction

• Sarcasm detection is the task of predicting whether a given piece of text is sarcastic.

• The ruling paradigm in sarcasm detection research is to design features that incorporate contextual information to understand context incongruity that lies at the heart of sarcasm.

• ‘I love being ignored’ : Incorporating context incongruity using sentiment flips.

• What happens in case of sentences with few or no sentiment words?
Motivation

A woman needs a man like a fish needs bicycle

Word2Vec similarity(man, woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

Can word embedding-based features when augmented to features reported in prior work improve the performance of sarcasm detection?
Word embedding based features

(Stop words removed)

**Unweighted similarity features (S):** For every word and word pair,
1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

**Distance-weighted similarity features (WS):** 4 S features weighted by linear distance between the two words

<table>
<thead>
<tr>
<th></th>
<th>man</th>
<th>woman</th>
<th>fish</th>
<th>needs</th>
<th>bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>-</td>
<td>0.766</td>
<td>0.151</td>
<td>0.078</td>
<td>0.229</td>
</tr>
<tr>
<td>woman</td>
<td>0.766</td>
<td>-</td>
<td>0.084</td>
<td>0.060</td>
<td>0.229</td>
</tr>
<tr>
<td>fish</td>
<td>0.151</td>
<td>0.084</td>
<td>-</td>
<td>0.022</td>
<td>0.130</td>
</tr>
<tr>
<td>needs</td>
<td>0.078</td>
<td>0.060</td>
<td>0.022</td>
<td>-</td>
<td>0.060</td>
</tr>
<tr>
<td>bicycle</td>
<td>0.229</td>
<td>0.229</td>
<td>0.130</td>
<td>0.060</td>
<td>-</td>
</tr>
</tbody>
</table>

**Both (S+WS):** 8 features
Experiment setup

• **Dataset**: 3629 Book snippets (759 sarcastic) downloaded from GoodReads website. Labeled by users with tags. We download ones with ‘sarcasm’ as sarcastic, ones with ‘philosophy’ as non-sarcastic.

• Five-fold cross-validation

• **Classifier**: SVM-Perf optimised for F-score

• **Configurations**:
  – Four prior works (augmented with our sets of features)
  – Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)
Observation: Only word embedding-based features will not suffice. ‘Augmentation’ to other known useful features necessary.
• **Observation**: Using word embedding-based features improves sarcasm detection, for multiple word embedding types and feature sets

![Table 3: Performance obtained on augmenting word embedding features to features from four prior works, for four word embeddings; L: Liebrecht et al. (2013), G: González-Ibáñez et al. (2011a), B: Buschmeier et al. (2014), J: Joshi et al. (2015)](image)
Conclusion

• Word embeddings can be used to design novel features for sarcasm detection

• Word embeddings do not operate well on their own as features

• When combined with past feature sets (based on punctuations, sentiment flips, affective lexicons, etc.), these word embedding-based features result in improved performance

• The performance is highest when Word2Vec embeddings are used (Several reasons: Large training corpus, Domain similarity, etc.)
Goal of NLP

• Science Goal
  – Understand human language behaviour

• Engineering goal
  – Unstructured Text $\rightarrow$ Structured Data
No “democracy”: Tail phenomenon and Language phenomenon

• Long tail Phenomenon: Probability is very low but not zero over a large number of phenomena.

• Language Phenomenon:
  – “people” which is predominantly tagged as “Noun” displays a long tail behaviour.
  – “laugh” is predominantly tagged as “Verb”.

21 Sept 2016  nlp-ml: fuss  124
Word embedding-based features for sarcasm detection
(To appear in EMNLP 2016)

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman, Are Word Embedding-based Features Useful for Sarcasm Detection?, EMNLP 2016, Austin, Texas, USA, November 1-5, 2016.
Introduction

• Sarcasm detection is the task of predicting whether a given piece of text is sarcastic

• The ruling paradigm in sarcasm detection research is to design features that incorporate contextual information to understand context incongruity that lies at the heart of sarcasm

• ‘I love being ignored’ : Incorporating context incongruity using sentiment flips

• What happens in case of sentences with few or no sentiment words?
Motivation

A **woman** needs a **man** like a **fish** needs **bicycle**

Word2Vec similarity(man,woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

*Can word embedding-based features when augmented to features reported in prior work improve the performance of sarcasm detection?*
Word embedding based features

(Stop words removed)

**Unweighted similarity features (S):** For every word and word pair,

1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

<table>
<thead>
<tr>
<th></th>
<th>man</th>
<th>woman</th>
<th>fish</th>
<th>needs</th>
<th>bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>-</td>
<td>0.766</td>
<td>0.151</td>
<td>0.078</td>
<td>0.229</td>
</tr>
<tr>
<td>woman</td>
<td>0.766</td>
<td>-</td>
<td>0.084</td>
<td>0.060</td>
<td>0.229</td>
</tr>
<tr>
<td>fish</td>
<td>0.151</td>
<td>0.084</td>
<td>-</td>
<td>0.022</td>
<td>0.130</td>
</tr>
<tr>
<td>needs</td>
<td>0.078</td>
<td>0.060</td>
<td>0.022</td>
<td>-</td>
<td>0.060</td>
</tr>
<tr>
<td>bicycle</td>
<td>0.229</td>
<td>0.229</td>
<td>0.130</td>
<td>0.060</td>
<td>-</td>
</tr>
</tbody>
</table>
Experiment setup

- **Dataset**: 3629 Book snippets (759 sarcastic) downloaded from GoodReads website. Labeled by users with tags. We download ones with ‘sarcasm’ as sarcastic, ones with ‘philosophy’ as non-sarcastic.

- Five-fold cross-validation

- **Classifier**: SVM-Perf optimised for F-score

- **Configurations**:
  - Four prior works (augmented with our sets of features)
  - Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)
Results (1/2)

<table>
<thead>
<tr>
<th>Features</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>67.2</td>
<td>78.8</td>
<td>72.53</td>
</tr>
<tr>
<td>S</td>
<td>64.6</td>
<td>75.2</td>
<td>69.49</td>
</tr>
<tr>
<td>WS</td>
<td>67.6</td>
<td>51.2</td>
<td>58.26</td>
</tr>
<tr>
<td>Both</td>
<td>67</td>
<td>52.8</td>
<td>59.05</td>
</tr>
</tbody>
</table>

**Observation:** Only word embedding-based features will not suffice. ‘Augmentation’ to other known useful features necessary.
Observation: Using word embedding-based features improves sarcasm detection, for multiple word embedding types and feature sets.
Conclusion

• Word embeddings can be used to design novel features for sarcasm detection

• Word embeddings do not operate well on their own as features

• When combined with past feature sets (based on punctuations, sentiment flips, affective lexicons, etc.), these word embedding-based features result in improved performance

• The performance is highest when Word2Vec embeddings are used (Several reasons: Large training corpus, Domain similarity, etc.)