NLP
Introduction to Natural Language Processing

Introduction
Language and Communication

- **Speaker**
  - Intention (goals, shared knowledge and beliefs)
  - Generation (tactical)
  - Synthesis (text or speech)

- **Listener**
  - Perception
  - Interpretation (syntactic, semantic, pragmatic)
  - Incorporation (internalization, understanding)

- **Both**
  - Context (grounding)
Basic NLP Pipeline

- (U)nderstanding and (G)eneration
Introduction to NLP

Examples of Text
What NLP is not about

- Romeo loves Juliet.
- ZZZZZZ is a great stock to buy.
What it is about (1/2)

• After the ball, in what is now called the "balcony scene", Romeo sneaks into the Capulet orchard and overhears Juliet at her window vowing her love to him in spite of her family's hatred of the Montagues.
What it is about (2/2)

- **ZZZZZ Resources** (NYSE:ZZZZZ) in their third quarter financials present a picture of a company with a relatively high amount of debt versus shareholder equity, and versus revenues. The company had total liabilities in the third quarter of $4,416 million versus shareholders' equity of only $1,518 million. That is a very high 3 to 1 debt to equity ratio. The company had third quarter revenues of $306 million. On an annualized basis, revenues would come out to $1,224 million. The company's debt level is almost 3 times its annual revenues. And remember, third quarter revenue is from before oil prices dropped in half. It looks like ZZZZZ may have bitten off more than it can chew.

- **XXXXX Petroleum** (NYSE:XXXXX) is another company whose third quarter financials present a relatively high debt load. The company had total liabilities in the third quarter of $3,272 million versus shareholder equity of only $1,520 million. That represents a high 2 to 1 debt to equity ratio. The company had third quarter revenues of $350 million. On an annualized basis revenues would come out to $1,400 million. The company's debt is more than 2 times its annual revenue. While XXXXX is a very good operator, it looks like they have taken on the high debt strategy at the wrong time.

- **YYYYY Energy** (NYSE:YYYYY) has a relatively high debt load according to their third quarter financials. The company had total liabilities of $2,026 million versus shareholder equity of $1,079. That is almost a 2 to 1 debt to equity ratio. Their third quarter revenues were $207 million. When annualized, their third quarter revenues come out to $827 million. The company's debt is almost 2 1/2 times its annualized revenues, and that is before the collapse of oil prices.
Brazil crowds attend funeral of late candidate Campos

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday. They attended a funeral Mass and filled the streets of the city of Recife to follow the passage of his coffin. Later this week, Mr Campos's Socialist Party is expected to appoint former Environment Minister Marina Silva as a replacement candidate.

Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo. Investigators are still trying to establish the exact causes of the crash, which killed six other people.

Mr Campos's private plane - a Cessna 560XL - was travelling from Rio de Janeiro to the sea-side resort of Guarujá, near the city of Santos.

President Dilma Rousseff, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.
Brazil crowds attend funeral of late candidate Campos

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday. They attended a funeral Mass and filled the streets of the city of Recife to follow the passage of his coffin. Later this week, Mr Campos's Socialist Party is expected to appoint former Environment Minister Marina Silva as a replacement candidate.

Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo. Investigators are still trying to establish the exact causes of the crash, which killed six other people. Mr Campos's private plane - a Cessna 560XL - was travelling from Rio de Janeiro to the sea-side resort of Guaruja, near the city of Santos.

President Dilma Rousseff, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.

Why did I highlight some of the phrases above?
Highlighted Phrases

• Brazil crowds attend funeral of late candidate Campos
  – Current event
• Mr Campos's jet crashed in bad weather in Santos
  – Background event
• Mr Campos's Socialist Party is expected to appoint…
  – Speculation
• President Dilma Rousseff
  – Property
• They attended a funeral Mass
  – Pronominal reference to an entity in a previous sentence
Genres of Text

- Blogs, emails, press releases, chats, debates, etc.
- Each presents different challenges to NLP
Turmoil in Peru as PM quits

LIMA, Peru (Reuters) - Peru's new Premier Víctor Raúl think Thursday he had saved all his personal misfortunes to spend his time in the country's political...
Induction of Influenza-Specific Mucosal Immunity by an Attenuated Recombinant Sendai Virus

Thuc-vy L. Le¹, Elena Mironova², Dominique Garcin², Richard W. Com dns¹*

¹Department of Microbiology and Immunology and Emory Vaccine Center, Emory University School of Medicine, Atlanta, Georgia, United States of America,
²Department of Microbiology and Molecular Medicine, University of Geneva School of Medicine, Geneva, Switzerland

Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,9]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. Sendai virus is a member of the Paramyxoviridae family, belongs in the genus respirivirus and shares 60–80% sequence homology to human parainfluenza virus type 1 (HPV-1) [9,10]. The viral genome consists of a negative sense, non-segmented RNA. Although Sendai virus was originally isolated from humans during an outbreak of pneumonia [11] subsequent human exposures to Sendai virus have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPV-1. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against HPV-1 [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and recombinant viruses have been reported to efficiently transduce luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells [16]. These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of parainfluenza virus infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

Figure 1. Generation of the GP42-H1 vector. A) Representation of wild-type Sendai virus gene construct (i) illustrating the major viral genes N, P, M, F, HN and L. The left promoter (P₄) and right promoter (P₅) function as the genomic promoter and anti-genomic promoter respectively. Seven gene boundaries that encode the conserved regulatory transcription start and transcription stop sequence are represented numerically. The mutant GP42-5eV (ii) genomic RNA is identical to WT Sendai virus with the exception of 3' P₅ in which 42 nucleotides of the P₅ were replaced with the corresponding sequence from P₄ [3,9]. Additional transcription start, stop, polyadenylation sequences and a unique Mu I restriction site were introduced into the intergenic region between the Sendai M and F genes. Using the unique Mu I restriction site, the GFP or HA gene was inserted (respecting the rule of si), generating the recombinant Sendai GP42-GFP [9,19] or GP42-H1 vectors (iii). B) Recombinant GP42-H1 virus was cultured in BSR-T7 cells for three passages. Cell free supernatant containing virus suspensions were collected and used to infect CV-1 cells (refer to Materials and Methods). Proteins from GP42-GFP or GP42-H1 infected cell extracts were resolved on SDS-PAGE and screened for GFP (left) or HA (right) expression by western analysis. Mock infected cells and allantoic fluid from PR/8 infected eggs are also shown.

doi:10.1371/journal.pone.0018780.g001

Plos ONE
DOI: 10.1371/journal.pone.0018780

Plos ONE
DOI: 10.1371/journal.pone.0018780
Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. **Sendai virus** is a member of the Paramyxoviridae family, belongs in the genus respirovirus and shares 60–80% sequence homology to **human parainfluenza virus type 1 (HPIV-1)** [9,10]. The viral genome consists of a negative sense, non-segmented RNA. Although **Sendai virus** was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to **Sendai virus** have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and **Sendai virus** infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made **Sendai virus** a mouse model for **HPIV-1**. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against **HPIV-1** [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. **Sendai virus** naturally infects the respiratory tract of mice and **recombinant viruses** have been reported to efficiently transduce **luciferase**, **lac Z** and **green fluorescent protein (GFP)** genes in the airways of **mice** or **ferrets** as well as **primary human nasal epithelial cells** [16].

These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of **parainfluenza virus** infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

<table>
<thead>
<tr>
<th>Named entities + variants</th>
<th><strong>human parainfluenza virus type, HPIV-1</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speculation</td>
<td>(reported, suggesting)</td>
</tr>
<tr>
<td>Species</td>
<td><strong>human</strong></td>
</tr>
<tr>
<td>Cell types</td>
<td>(nasal epithelial cells)</td>
</tr>
<tr>
<td>Facts</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td></td>
</tr>
</tbody>
</table>
Medical Records

TITLE: ACUTE CARE VISIT
DATE OF NOTE: FEB 04, 2000@11:18
ENTRY DATE: FEB 04, 2000@11:20
AUTHOR: EXP COSIGNER:
URGENCY: STATUS: COMPLETED

Chief Complaint: Patient notes 1 month history of blurred vision and frequent urination

HISTORY OF PRESENT ILLNESS:

DEMO, FATHER is a 44 year-old MALE who presents complaining of blurred vision for the past 1 month. He finds it is difficult for him to read clearly and is even effecting his driving. He also notes that he has been getting up to the bathroom frequently, esp. at night. He now routinely urinates 3-4 times a night. He is not aware of any particular weight loss, but does feel thirsty much of the time.

PAST MEDICAL HISTORY:
Illnesses: Hypertension

Surgeries: None

Allergies: PENICILLINS

Medications:
1) HYDROCHLOROTHIAZIDE 25MG TAB** Qty: 45 ACTIVE
   for 90 days Sig: TAKE ONE-HALF TABLET Refills: 0
   MOUTH EVERY MORNING THC BLOOD PRESSURE
2) METOPROLOL 25MG XL TAB Qty: 90 for 90 ACTIVE
   days Sig: TAKE ONE TABLET MOUTH 0DAY Refills: 0
   FOR THE HEART

FAMILY HISTORY:
Diabetes Father, Sibling, Grandparent

Literary Texts

• Project Gutenberg (http://www.gutenberg.org/browse/scores/top)

• A team of horses passed from Finglas with toiling plodding tread, dragging through the funereal silence a creaking waggon on which lay a granite block. The waggoner marching at their head saluted.

• There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question.

• Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife. Their house was small, for the lumber to build it had to be carried by wagon many miles. There were four walls, a floor and a roof, which made one room; and this room contained a rusty looking cookstove, a cupboard for the dishes, a table, three or four chairs, and the beds. Uncle Henry and Aunt Em had a big bed in one corner, and Dorothy a little bed in another corner. There was no garret at all, and no cellar--except a small hole dug in the ground, called a cyclone cellar, where the family could go in case one of those great whirlwinds arose, mighty enough to crush any building in its path. It was reached by a trap door in the middle of the floor, from which a ladder led down into the small, dark hole.
Try parsing this

“Bloat is one of the co-tenants of the place, a maisonette erected last century, not far from the Chelsea Embankment, by Corydon Throsp, an acquaintance of the Rossettis' who wore hair smocks and liked to cultivate pharmaceutical plants up on the roof (a tradition young Osbie Feel has lately revived), a few of them hardy enough to survive fogs and frosts, but most returning, as fragments of peculiar alkaloids, to rooftop earth, along with manure from a trio of prize Wessex Saddleback sows quartered there by Throsp's successor, and dead leaves off many decorative trees transplanted to the roof by later tenants, and the odd unstomachable meal thrown or vomited there by this or that sensitive epicurean—all got scumbled together, eventually, by the knives of the seasons, to an impasto, feet thick, of unbelievable black topsoil in which anything could grow, not the least being bananas.”

Do you know the source?
Quiz Answer

• “Gravity’s Rainbow” (by Thomas Pynchon), known for its use of very arcane words and complicated sentence (and plot) structure.
• Another such work is “Finnegans Wake” by James Joyce.
• Poetry is even more difficult.
Introduction to NLP

Funny Sentences
Silly Sentences

- Children make delicious snacks
- Stolen painting found by tree
- I saw the Rockies flying to San Francisco
- Court to try shooting defendant
- Ban on nude dancing on Governor’s desk
- Red tape holds up new bridges
- Government head seeks arms
- Cameron wins on budget, more lies ahead
- Local high school dropouts cut in half
- Hospitals are sued by seven foot doctors
- Dead expected to rise
- Miners refuse to work after death
- Patient at death's door – doctors pull him through
- In America a woman has a baby every 15 minutes. How does she do that?
### Ambiguous Recommendations

<table>
<thead>
<tr>
<th>Lexical ambiguity</th>
<th>For a chronically absent employee</th>
<th>A man like him is hard to find.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Ambiguous Recommendations

<table>
<thead>
<tr>
<th>Lexical ambiguity</th>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>For a chronically absent employee</td>
<td>A man like him is hard to find.</td>
</tr>
<tr>
<td>For a dishonest employee</td>
<td>He's an unbelievable worker.</td>
</tr>
<tr>
<td>For a lazy employee</td>
<td>You would indeed be fortunate to get this person to work for you.</td>
</tr>
<tr>
<td>For the office drunk</td>
<td>Every hour with him was a happy hour.</td>
</tr>
</tbody>
</table>
# Ambiguous Recommendations

## Lexical ambiguity

<table>
<thead>
<tr>
<th>For a chronically absent employee</th>
<th>A man like him is hard to find.</th>
</tr>
</thead>
<tbody>
<tr>
<td>For a dishonest employee</td>
<td>He's an unbelievable worker.</td>
</tr>
<tr>
<td>For a lazy employee</td>
<td>You would indeed be fortunate to get this person to work for you.</td>
</tr>
<tr>
<td>For the office drunk</td>
<td>Every hour with him was a happy hour.</td>
</tr>
</tbody>
</table>

## Structural ambiguity

<table>
<thead>
<tr>
<th>For a chronically absent employee</th>
<th>It seemed her career was just taking off.</th>
</tr>
</thead>
<tbody>
<tr>
<td>For a dishonest employee</td>
<td>Her true ability was deceiving.</td>
</tr>
<tr>
<td>For a stupid employee</td>
<td>I most enthusiastically recommend this candidate with no qualifications whatsoever.</td>
</tr>
<tr>
<td>For the office drunk</td>
<td>He generally found him loaded with work to do.</td>
</tr>
</tbody>
</table>

## Scope ambiguity

<table>
<thead>
<tr>
<th>For an employee who is not worth further consideration as a job candidate</th>
<th>All in all, I cannot say enough good things about this candidate or recommend him too highly.</th>
</tr>
</thead>
<tbody>
<tr>
<td>For an employee who is so unproductive that the job is better left unfilled</td>
<td>I can assure you that no person would be better for the job.</td>
</tr>
</tbody>
</table>

## Other

<table>
<thead>
<tr>
<th>For a lazy employee</th>
<th>He could not care less about the number of hours he has to put in.</th>
</tr>
</thead>
<tbody>
<tr>
<td>For an employee who is not worth further consideration as a job candidate</td>
<td>I would urge you to waste no time in making this candidate an offer of employment.</td>
</tr>
<tr>
<td>For a stupid employee</td>
<td>There is nothing you can teach a man like him.</td>
</tr>
</tbody>
</table>

From Beatrice Santorini’s collection