Discovery of Linguistic Relations Using Lexical Attraction

Deniz Yuret
Overview

• Motivation

• Demonstration

• Theory, Learning, Algorithm

• Evaluation

• Contributions
Syntax and Semantics independently constrain linguistic relations

- I saw the Statue of Liberty flying over New York.
  – Lenat, 1984

- I hit the boy with the girl with long hair with a hammer with vengeance.
  – Schank, 1973

- Colorless green ideas sleep furiously.
  – Chomsky, 1956
Contributions of this thesis

- Opening a door for the use of common sense knowledge in language processing and acquisition.

- A learning paradigm that bootstraps by interdigitating learning with processing.
Bringing common sense into language

John eats ice-cream

- John
- ice-cream
- eat

S (Subject)
O (Object)
Bootstrapping by interdigitating learning and processing
Phrase structure versus dependency structure

The glorious sun will shine in the winter

The glorious sun will shine in the winter
Discovery of Linguistic Relations
An Example

Simple Sentence 1/5
(Before training)

* these people also want more government money for education. *
* these people also want more government money for education. *
these people also want more government money for education.
these people also want more government money for education.
* these people also want more government money for education. *
Bringing common sense into language
The theory

John
ice-cream
eat

John eats ice-cream
S
O
A Theory of Syntactic Relations

- Lexical attraction is the likelihood of a syntactic relation

- The context of a word is given by its syntactic relations

- Syntactic relations can be formalized as a graph

- Entropy is determined by syntactic relations
\[ H = - \sum p_i \log p_i \]

The information content of a word:

<table>
<thead>
<tr>
<th>Word</th>
<th>Information Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>The IRA is fighting British rule in Northern Ireland</td>
<td>99.28 bits</td>
</tr>
</tbody>
</table>

Total: 99.28 bits
The word pair and relative information:

<table>
<thead>
<tr>
<th>Northern</th>
<th>Ireland</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.60</td>
<td>14.65</td>
</tr>
</tbody>
</table>

1.48

<table>
<thead>
<tr>
<th>Northern</th>
<th>Ireland</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.60</td>
<td>3.53</td>
</tr>
</tbody>
</table>
The lexical attraction link:

Northern \hspace{1cm} Ireland
12.60 \hspace{1cm} 14.65

11.12
Language Model Determines the Context

The IRA is fighting British rule in Northern Ireland

4.20 12.90 3.73 10.54 8.66 5.96 3.57 9.25 3.53

Total: 99.28 → 62.34 bits
Context should be determined by syntactic relations:

The man with the dog spoke

The man with the dog spoke
Context should be determined by syntactic relations:

The IRA is fighting British rule in Northern Ireland

Total: 62.34 $\rightarrow$ 49.85 bits
Dependency structure is acyclic:

- Mathematically: cannot use all the lexical attraction links in a cycle.

- Linguistically: cannot construct a consistent head-modifier structure.
Syntactic relations form a planar tree:
(Links do not cross)

I met the woman in the red dress in the afternoon

? I met the woman in the afternoon in the red dress
Syntactic relations form a planar tree:
(Links do not cross)

- Hays and Lecerf (1960) discovered that (almost) all sentences in a language are planar.

- Gaifman (1965) proved that a planar dependency grammar can generate the same set of languages as a context free grammar.

- Planar trees can be encoded with constant number of bits per word.
Cayley’s formula for counting trees:

\[ T(n) = n^{n-2} \]

Planar trees are polynomial in \( n \):

The IRA is fighting British rule in Northern Ireland

Encoding: LPLLPPRRLPRLPLPPP
L:10 R:11 P:0
Upper bound: 3 bits per word
Lexical attraction is symmetric.

The IRA is fighting British rule.
Lexical attraction is symmetric

\[ S = (W, L, w_0) \]
\[ W = \{ w_i \} \]
\[ L = \{ (w_i, w_j) \} \]

\[
P(S) = P(L)P(w_0) \prod_{(w_i, w_j) \in L} P(w_j \mid w_i)
\]
\[
= P(L)P(w_0) \prod_{(w_i, w_j) \in L} \frac{P(w_i, w_j)}{P(w_i)}
\]
\[
= P(L) \prod_{w_i \in W} P(w_i) \prod_{(w_i, w_j) \in L} \frac{P(w_i, w_j)}{P(w_i)P(w_j)}
\]
Dependency structure is an undirected, acyclic, planar graph:

The IRA is fighting British rule in Northern Ireland
Information in a Sentence =

Information in Words

+ Information in the Tree

- Mutual Information in Syntactic Relations
The Memory
The memory observes the processor

- kick the ball now
- kick the ball now
- kick the ball now
Learning simple structures

1. Kick the ball now
2. Throw the ball at
3. With the ball in
4. Kick the ball now
Simple structures help see complex structures.
Learning complex structures
• We need to discover the best linkage.

* these people also want more government money for education. *
• Words are read in left to right order.

\[ \text{\textit{these}} \]
• New word considers links with previous words.

348
* these people
118
• Cycles are not allowed.

• Link with minimum score gets rejected.

\[\begin{align*}
\text{55} & \quad & \text{**these people**} \\
\begin{array}{c}
118 \\
348
\end{array}
\end{align*}\]
• Link with negative value not accepted.

\begin{itemize}
\item \[-164\]
\item these people also
\item \(\{118, 348\}\)
\end{itemize}
• Link crossing not allowed.

• Link with minimum score gets eliminated.

* these people also want

118 348 178 143 315
* these people also want

118  348  315  143  261
• The two constraints straighten out previous mistakes by eliminating bad links.

* these people also want more government money
• Eliminating bad links 2/3

* these people also want more government money
• Eliminating bad links 3/3

* these people also want more government money

118 348 315 143 66 401 43 209
- New link can knock off old link in cycle.

* these people also want more government money for education

```
118  348  315  143  401  43  209  261  258  392
```
• The final result.

* these people also want more government money for education.
Discovery of Linguistic Relations
Using Lexical Attraction

A demonstration

- Long distance link
- Complex noun phrase
- Syntactic ambiguity
the cause of his death friday was not given.
* the cause of his death friday was not given. *
* the cause of his death friday was not given. *
Complex Noun Phrase 1/4
(After 10,000 words of training)

* the new york stock exchange composite index fell. *
Complex Noun Phrase 2/4
(After 100,000 words of training)

* the new york stock exchange composite index fell . *
Complex Noun Phrase 3/4
(After 1,000,000 words of training)

* the new york stock exchange composite index fell . *
the new york stock exchange composite index fell.
Syntactic Ambiguity 1/3
(After 1,000,000 words of training)

* many people died in the clashes in the west in september. *
many people died in the clashes in the west in september.
Syntactic Ambiguity 2/3
(After 500,000 words of training)

* a number of people protested . *

* the number of people increased . *
Syntactic Ambiguity 2/3
(After 5,000,000 words of training)

* a number of people protested . *

* the number of people increased . *
* the driver saw the airplane flying over washington . *

* the pilot saw the train flying over washington . *
Syntactic Ambiguity 3/3
(After 10,000,000 words of training)

* the driver saw the airplane flying over washington . *

* the pilot saw the train flying over washington . *
Results

- Evaluation criteria
- Upper and lower bounds
- Link accuracy
- Related work
Evaluation criteria: Content-word links

I saw the mountains flying over New York

People want more money for education
Training

- Up to 100 million words of Associated Press material.

Testing

- 200 out-of-sample sentences.

- Selected from 5000 word vocabulary (90% of all the words seen in the corpus).

- 3152 words (15.76 words per sentence).

- Hand parsed with 1287 content-word links.
Accuracy:

\( n_1 = \text{human links} \)
\( n_2 = \text{program links} \)
\( n_{12} = \text{common links} \)

- Precision = \( \frac{n_{12}}{n_2} \)
- Recall = \( \frac{n_{12}}{n_1} \)
Lower bound:

Random lexical attraction → 8.9% precision, 5.4% recall

Linking every adjacent word → 41% recall

Upper bound:

85% of syntactically related pairs have positive lexical attraction
Recording adjacent pairs

![Graph showing Procedure 1: Recording adjacent pairs]

Precision = 67%
Recall = 41%
Procedure 2: Recording all pairs

Precision = 55%
Recall = 48%
Using feedback from processor

Procedure 3: Recording pairs selected by processor

Precision = 62%
Recall = 52%
Related work

- Magerman and Marcus, 1990
- Lari and Young, 1990
- Pereira and Schabes, 1992
- Briscoe and Waegner, 1992
- Carroll and Charniak, 1992
- Stolcke, 1994
- Chen, 1996
- de Marcken, 1996
de Marcken, 1995

\[
\begin{align*}
S & \rightarrow CP \\
CP & \rightarrow AP \ C \\
AP & \rightarrow A \ BP \\
BP & \rightarrow B \\
CP & \rightarrow AP \ C \\
BP & \rightarrow \ AP \ B \\
CP & \rightarrow BP \ C
\end{align*}
\]
Lessons learned

- Training with words instead of parts of speech enable the program to learn common but idiosyncratic usages of words.

- Not committing to early generalizations prevent the program from making irrecoverable mistakes early.

- Using a representation that makes the relevant features (such as syntactic relations) explicit simplifies learning.
Contributions

• Opening a door for common sense in language

• Bootstrapping from zero by interdigitating learning and processing
Future Work

- Second degree models
- History mechanism
- Categorization and generalization