Modeling ESL Word Choice Similarities By Representing Word Intensions and Extensions

Huichao Xue, Rebecca Hwa
Department of Computer Science
University of Pittsburgh
Outline

• Task – Building Confusion Sets for Grammatical Error Correction (GEC) systems
• Idea – Simulating ESL Word Learning
  – Words have *Intensions* and *Extensions*
  – We simulate learning both components
• Simulation Models
• Experimental Results
Grammatical Error Correction (GEC)

• English as Second Language (ESL) learners make grammar errors.

ESL (English as Second Language speakers) TEXT

... Beside the factor of the lead user, ...
... Besides With the help modern technology, ...
... of ...

• We try to build systems that automatically detect and correct these mistakes.
Confusion Sets in GEC systems

- Grammatical Error Correction by Classification

Which is Correct?

beside about along besides around

- Confusion sets need to be pre-defined
Ways to build Confusion Sets?

Hiring human experts to filter the initially large sets (Liu, 2010)

Count from annotated ESL corpora (Rozovskaya, 2010)

- to → for: 35 times
- for → of
- to → at

Is there a better (cheaper) way?
Proposal: Building Confusion Sets from Normal English Corpora

• Normal English corpora is enough

• Our idea:
  – Simulate how learners learn English words.
  – Find out which words are similar

• Test our idea on prepositions

36 most frequent prepositions
about, along, among, around, as, at, beside, besides, between, by, down, during, except, for, from, in, inside, into, of, off, on, onto, outside, over, through, to, toward, towards, under, underneath, until, up, upon, with, within, without.
Knowledge of a word – Intension and Extension

What do people learn about words?

Intension

Any property or quality connoted by a word.

Extension

Things the word applies to. Characterized by contexts it appears in.

Example:

1. used as a function word to indicate a point of reckoning
   <north of the lake>

2. a used as a function word to indicate origin or derivation
   <a man of noble birth>

b used as a function word to indicate the cause, motive, or reason
<died of flu>

c: by <plays of Shakespeare>
Three Views on Word Learning

1. Concentrating on intensions
   – E.g. lexical semantics
   – Words with similar meanings are confusabe

2. Concentrating on extensions
   – E.g. language modeling
   – Words occur in similar contexts are confusabe

3. Our view – both intensions and extensions
   – Ultimately, compatible intension and extension
Outline

• Task – Building Confusion Sets for Grammatical Error Correction (GEC) systems
• Idea – Simulating ESL Word Learning

• Simulation Models
  – Intension only
  – Extension only
  – Both (proposed model)

• Experimental Results
Intension Based Model

• Learners’ main goal is to learn about words’ meanings.
  Confusion is caused by meanings’ similarity.

• Meanings are learned from reading text. (Fischer, 1990)
  – We simulate using Distributional Models (Pereira et al., 1993; Lee, 1999)
    • The similarities are shown to correlate with word meaning similarities
  – We then fill in confusion sets with the most similar words
Intension Based Model
(distributional models)

Dimensions are contexts (very many):

|    | __ that | __ you | __ course | ...
|----|---------|--------|-----------|-------
| of | 0.50    | 0.25   | 1.00      | ...
| to | 0.0     | 0.50   | 0.00      | ...
| for| 0.50    | 0.25   | 0.00      | ...

Words are Vectors of Distributions: Pr(word|context)

Normal Metrics for Similarities:
1. Euclidean Distance
2. KL-Divergence
3. Cosine Similarities
Intension Based Model

Corpus
... need of ...
... need for ...
... of you ...
... to you ...
... for you ...
... of course ...
...

<table>
<thead>
<tr>
<th></th>
<th>of</th>
<th>to</th>
<th>for</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>to</td>
<td>1.5</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>for</td>
<td>2.0</td>
<td>2.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Distributional vectors

of → 0.50 0.25 1.00 ...
to → 0.0 0.50 0.00 ...
for → 0.50 0.25 0.00 ...

Calculate distance
Pick the closest

Word | Confusion set
-----|------------------
of   | of, to
for  | for, of
Outline

• Task – Building Confusion Sets for Grammatical Error Correction (GEC) systems
• Idea – Simulating ESL Word Learning

• Simulation Models
  – Intension Only
  – Extension Only
  – Both (Proposed model)

• Experimental Results
Extension Based Model (Word Selector)

- Learners’ main goal is to learn how to choose words.
- We use classifiers to simulate learners
  Rerun the trained classifier on training data to collect mistakes
# Extension Based Model

## Corpus

| ... need of ... |
| ... need for ... |
| ... of you ... |
| ... to you ... |
| ... to you ... |
| ... for you ... |
| ... of course ... |
| ... |

## Word Selector

<table>
<thead>
<tr>
<th>word</th>
<th>true</th>
</tr>
</thead>
<tbody>
<tr>
<td>need</td>
<td>of __⇒ of</td>
</tr>
<tr>
<td>you</td>
<td>__⇒ to</td>
</tr>
<tr>
<td>course</td>
<td>__⇒ of</td>
</tr>
<tr>
<td>...</td>
<td>...⇒ ...</td>
</tr>
</tbody>
</table>

## Word

<table>
<thead>
<tr>
<th>Word</th>
<th>Confusion set</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>of, to</td>
</tr>
<tr>
<td>to</td>
<td>to, for</td>
</tr>
<tr>
<td>for</td>
<td>for, to</td>
</tr>
</tbody>
</table>
Outline

• Task – Building Confusion Sets for Grammatical Error Correction (GEC) systems
• Idea – Simulating ESL Word Learning
• Simulation Models
  – Intension Only
  – Extension Only
  – Both (Proposed model)
• Experimental Results
Proposed method: Combine Intension and Extension

Word usages (extension) involve the understandings of words’ meanings (intension).

Example:
- country’s needs __ water
- 1. water is the object
- 2. water is necessary for the country

√ of to ...

Extension

Context Step 1 Intension Step 2 Word Choice
Combining Intension and Extension
Mathematical Formulation

- Intensions are represented as points in an Euclidean space $S$:
  - Prepositions are vectors in $S$: $v_{of}, v_{to}, \ldots$
  - We assume $S = R^{36}$

- The first step of word choice decision is formalized as $f: Context \rightarrow S$

  e.g. Need __ $\rightarrow [1.2, 3.5, -1.2, \ldots, 1]^T$
Combining Intension and Extension
Learning Simulation

Corpus
... need of ...
... need for ...
... of you ...
... to you ...
... to you ...
... for you ...
... of course ...
...

\[
\min_{f, \text{of}, \text{for}, \ldots} \left\| f(\text{need } \_ ) - \text{of} \right\|^2 + \ldots \quad \text{s.t. Area}(\text{of, for, } \ldots) \geq 1
\]

Optimal \( \text{of, for, } \ldots \) indicate the ultimate understandings of prepositions’ intensions

Relevance Component Analysis (RCA) (Bar et al., 2006) solves this optimization problem
Combining Intension and Extension (RCA)

**Corpus**

- ... need of ...
- ... need for ...
- ... of you ...
- ... to you ...
- ... to you ...
- ... for you ...
- ... of course ...
- ...

<table>
<thead>
<tr>
<th></th>
<th>of</th>
<th>to</th>
<th>for</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>0</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>to</td>
<td>2.0</td>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>for</td>
<td>1.5</td>
<td>2.5</td>
<td>0</td>
</tr>
</tbody>
</table>

**RCA**

Calculate distance

Pick the closest

**Word**

- of
- to
- for

**Confusion set**

- of, for
- to, of
- for, of
Outline

• Task – Building Confusion Sets for Grammatical Error Correction (GEC) systems
• Idea – Simulating ESL Word Learning
  – Words have *Intensions* and *Extensions*
  – We simulate learning both components
• Simulation Models
• Experimental Results
Experimental Setup – task

• Constructing confusion sets for the 36 most frequent prepositions

• Evaluate extrinsically – end-to-end GEC system
  – On NUCLE corpus (Chinese ESL speakers)
  – filter training samples using the confusion set (Rozovskaya et al., 2010)
  – Evaluate by $F_1$ score
Experimental Setup – Confusion Set Construction Methods

• The three automatic methods
  • Intension only – kl div, euc dist, cos sim
  • Extension only – preposition selector
  • Combined – RCA
• all preps: containing all prepositions
• gold: containing the most frequently confused words according to real ESL corpus (NUCLE).
Experimental Setup – data and features

- FBIS corpus, GIZA++, Stanford Parser
- Indicative contextual features (Tetreault et al 2010):

  我 在 周日 去 得 教堂 。

  I went to the church on Sunday.

- **Gov**: dependency governors of the preposition
- **Obj**: dependency objects of the preposition
- **GovTag**: part-of-speech tags of Gov
- **ObjTag**: part-of-speech tags of Obj
- **L1-Trans**: L1 translation of the preposition

We incorporate **L1-Trans** to consider the L1 background of speakers.
Experiments – Questions

• Our proposed method V.S using annotated ESL corpus
• Intension/Extension only V.S Intension and Extension combined
• How do features affect our methods?
1. Automatic method is competitive
2. Learning intension and extension together is helpful
A Deeper Comparison

Precision
RCA and euc dist are more helpful

Coverage (Intrinsic Evaluation)
Proportion of ESL mistakes covered by the confusion sets
RCA and selector are better
Feature Sets and Confusion Sets’ Quality

**RCA method’s $F_1$ scores**

**Selector method’s $F_1$ scores**

RCA is more stable w.r.t. the feature set changes.
Conclusions

• A cheap method to automatically construct confusion sets
  – Simulation on normal English text
• The resulting confusion sets:
  – are able to improve GEC systems’ performance
  – Correlate well with real ESL mistakes
• Benefits from modeling interaction between intensional and extensional knowledge
Thank you!