Experiments with Three Approaches to Recognizing Lexical Entailment

Peter Turney and Saif Mohammad

2013
Outline of talk

- introduction
- semantic relations and lexical entailment
- related work
- performance measures
- three approaches to lexical entailment
- three datasets for lexical entailment
- experiments
- discussion of results
- limitations and future work
- conclusion
Introduction
Introduction – RTE

• RTE – Recognizing Textual Entailment
• sample problem from first RTE Challenge in 2005:
  • Text:
    – iTunes software has seen strong sales in Europe.
  • Hypothesis:
    – Strong sales for iTunes in Europe.
• does Text entail Hypothesis? Yes or no?
• generic NLP task with applications in text summarization, information retrieval, information extraction, question answering, machine translation, paraphrasing, …
Introduction – RTE and RLE

- to recognize entailment between sentences, must first recognize entailment between words:
  - Text:
    - George was bitten by a dog.
  - Hypothesis:
    - George was attacked by an animal.
  - bitten entails attacked
  - dog entails animal
  - RTE – Recognizing Textual Entailment requires
    RLE – Recognizing Lexical Entailment
Introduction – VSM

- will look at three approaches to RLE
- all three use Vector Space Model of Semantics
- core idea of VSM is distributional hypothesis
  - Distributional hypothesis: Words that occur in similar contexts tend to have similar meanings.
- represent a word by a vector of the contexts in which the word appears
- calculate entailment from vectors
Introduction – balAPinc

- first approach: balAPinc
  - (Kotlerman, Dagan, Szpektor, and Zhitomirsky-Geffet 2010)
- based on context inclusion hypothesis
  - Context inclusion hypothesis: If a word $a$ tends to occur in a subset of the contexts in which a word $b$ occurs ($b$ contextually includes $a$), then $a$ (the narrower term) tends to entail $b$ (the broader term).
  - inspired by formal logic, where ‘$a$ entails $b$’ means ‘the models in which $b$ is true include the models in which $a$ is true’
    - the models of $a$ are a subset of the models of $b$
Introduction – ConVecs

- second approach: ConVecs
  - (Baroni, Bernardi, Do, and Shan 2012)
- based on the context combination hypothesis
  - Context combination hypothesis: The tendency of \( a \) to entail \( b \) is correlated with some learnable function of the contexts in which \( a \) occurs and the contexts in which \( b \) occurs; some combinations of contexts tend to block entailment and others tend to allow entailment.
- hypothesis implies that the concatenation of the context vectors for \( a \) and \( b \) is suitable for supervised machine learning of lexical entailment
  - \(<a_1, a_2, ..., a_n, b_1, b_2, ..., b_n>\)
Introduction – SimDiffs

• third approach: SimDiffs
  • (Turney and Mohammad, under review)
• based on similarity differences hypothesis
  • Similarity differences hypothesis: The tendency of $a$ to entail $b$ is correlated with some learnable function of the differences in their similarities, $\text{sim}(a, r) - \text{sim}(b, r)$, to a set of reference words, $r \in R$; some differences tend to block entailment and others tend to allow entailment.

• consider ‘dog entails animal’ versus ‘table entails animal’, using $r = \text{life}$
  • $\text{sim}(\text{dog, life}) \approx \text{sim}(\text{animal, life})$
  • $\text{sim}(\text{table, life}) \neq \text{sim}(\text{animal, life})$
Semantic Relations and Lexical Entailment
Semantic Relations and Lexical Entailment

- balAPinc inspired by asymmetric similarity measures
- ConVecs and SimDiffs inspired by research on supervised learning of semantic relations
- Zhitomirsky-Geffet and Dagan (2009) argue lexical entailment does not correspond to classical relations
  - for example, some part-whole relations involve entailment and others don’t
- but a fine-grained semantic relation taxonomy can resolve this issue by distinguishing subcategories
  - Bejar, Chaffin, and Embretson (1991) have 79 subcategories of semantic relations
- entailment crosses coarse boundaries, but not fine
Related Work
Related Work

• asymmetric similarity measures for context vectors

• supervised learning of semantic relation classes
  • SemEval-2007 Task 4: Classification of Semantic Relations between Nominals (Girju et al. 2007)
  • SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals (Hendrickx, Kim, Kozareva, Nakov, Seaghdha, Pado, Pennacchiotti, Romano, and Szpakowicz 2010)
  • SemEval-2012 Task 2: Measuring Degrees of Relational Similarity (Jurgens et al. 2012)
Related Work

- only two papers approaching lexical entailment using ideas from semantic relation classification
  - Akhmatova and Dras (2009)
  - Baroni, Bernardi, Do, and Shan (2012)
Performance Measures
Performance Measures

- asymmetric similarity measures generate real-valued outputs
  - \( \text{sim}(a,b) \) is in \( R \)
    - balAPinc
- supervised learning of semantic relations yields binary-valued classes
  - word pair \(<a,b>\) is in class 0 (does not entail) or 1 (entails)
    - ConVecs, SimDiffs
- how to compare balAPinc with ConVecs and SimDiffs?
  - use learning algorithm that generates probabilities
  - apply threshold on balAPinc output
Performance Measures

- performance of real-valued scores:
  - AP0 – average precision with respect to class 0
  - AP1 – average precision with respect to class 1

- performance of binary-valued classifications:
  - precision
  - recall
  - F-measure
  - accuracy
Three Approaches
Three Approaches – balAPinc

• first approach: balAPinc
  • (Kotlerman, Dagan, Szpektor, and Zhitomirsky-Geffet 2010)

• based on context inclusion hypothesis
  • Context inclusion hypothesis: If a word $a$ tends to occur in a subset of the contexts in which a word $b$ occurs ($b$ contextually includes $a$), then $a$ (the narrower term) tends to entail $b$ (the broader term).

• asymmetric real-valued similarity measure
Three Approaches – balAPinc

\[
\text{rel}(f, F_w) = \begin{cases} 
1 - \frac{\text{rank}(f, F_w)}{|F_w|+1} & \text{if } f \in F_w \\
0 & \text{if } f \notin F_w
\end{cases}
\]

\[
\text{inc}(r, F_u, F_v) = \{ f \mid \text{rank}(f, F_u) \leq r \text{ and } f \in (F_u \cap F_v) \}
\]

\[
P(r, F_u, F_v) = \frac{|\text{inc}(r, F_u, F_v)|}{r}
\]

\[
\text{APinc}(u, v) = \frac{\sum_{r=1}^{|F_u|} [P(r, F_u, F_v) \cdot \text{rel}(f_{ur}, F_v)]}{|F_u|}
\]

\[
\text{LIN}(u, v) = \frac{\sum_{f \in F_u \cap F_v} [w_u(f) + w_v(f)]}{\sum_{f \in F_u} w_u(f) + \sum_{f \in F_v} w_v(f)}
\]

\[
\text{balAPinc}(u, v) = \sqrt{\text{APinc}(u, v) \cdot \text{LIN}(u, v)}
\]
Three Approaches – ConVecs

- second approach: ConVecs
  - (Baroni, Bernardi, Do, and Shan 2012)
- based on the context combination hypothesis
  - Context combination hypothesis: The tendency of \( a \) to entail \( b \) is correlated with some learnable function of the contexts in which \( a \) occurs and the contexts in which \( b \) occurs; some combinations of contexts tend to block entailment and others tend to allow entailment.
- supervised machine learning from labeled word pairs
- binary-valued class output
Three Approaches – ConVecs

- word pair \(<a,b>\) with label \{0,1\}
- represent \(<a,b>\) with feature vector
- context vectors for words \(a\) and \(b\):
  - \(<a_1, a_2, \ldots, a_n>\)
  - \(<b_1, b_2, \ldots, b_n>\)
- feature vector for machine learning:
  - \(<a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n>\)
- apply support vector machine (SVM) learning algorithm to labeled training data
Three Approaches – SimDiffs

- third approach: SimDiffs
  - (Turney and Mohammad, under review)
- based on similarity differences hypothesis
  - Similarity differences hypothesis: The tendency of $a$ to entail $b$ is correlated with some learnable function of the differences in their similarities, $\text{sim}(a, r) - \text{sim}(b, r)$, to a set of reference words, $r \in R$; some differences tend to block entailment and others tend to allow entailment.
- supervised machine learning from labeled word pairs
- binary-valued class output
Three Approaches – SimDiffs

- supervised machine learning, four sets of features
- \( \text{sim}_d = \) similarity in domain space
- \( \text{sim}_f = \) similarity in function space

\[
\begin{align*}
S_1 &= \{\text{sim}_d(a, r) - \text{sim}_d(b, r) \mid r \in R\} \\
S_2 &= \{\text{sim}_f(a, r) - \text{sim}_f(b, r) \mid r \in R\} \\
S_3 &= \{\text{sim}_d(a, r) - \text{sim}_f(b, r) \mid r \in R\} \\
S_4 &= \{\text{sim}_f(a, r) - \text{sim}_d(b, r) \mid r \in R\}
\end{align*}
\]

- \( R = 2,086 \) words from Basic English (Ogden 1930)
Three Datasets
Three Datasets – KDSZ dataset

- (Kotlerman, Dagan, Szpektor, and Zhitomirsky-Geffet 2010)
- dataset contains 3,772 word pairs
  - 1,068 labeled *entails*
  - 2,704 labeled *does not entail*
- three judges labeled the pairs
  - inter-annotator agreement between any two of the three judges varying from 90.0% to 93.5%
- pairs belong to a variety of semantic relation classes
  - dataset not designed with any attention to semantic relation classes
Three Datasets – BBDS dataset

• (Baroni, Bernardi, Do, and Shan 2012)
• dataset contains 2,770 word pairs
  • 1,385 labeled entails
  • 1,385 labeled does not entail
• all pairs labeled entails are hyponym–hypernym noun–noun pairs
  • such as 'pope entails leader'
• pairs were generated automatically from WordNet and then validated manually
Three Datasets – JMTH dataset

- (Jurgens, Mohammad, Turney, and Holyoak 2012)
- SemEval-2012 Task 2 dataset
  - contains 3,218 word pairs
  - labeled with 79 types of semantic relations
    - part-whole, cause-effect, sign-referent, ...
  - we convert the dataset into 2,308 word pairs
    - 1,154 labeled *entails*
    - 1,154 labeled *does not entail*
- manually created a mapping table to map 79 semantic relation types to binary *entails / does not entail*
  - 'hyponym entails hypernym'
  - 'cause entails effect'
Three Experiments
Experiments with the JMTH dataset

• split the dataset into three (roughly) equal parts
  • two development sets (Dev1 and Dev2)
  • one test set (Test)
• used Dev1 and Dev2 to tune parameters
  • balAPinc has threshold parameter $T$ to convert real-valued similarity score to binary-valued class
  • ConVecs has two parameters to tune for SVD, $k$ and $p$
  • SimDiffs has four parameters to tune for SVD, $k$ and $p$ for domain space and $k$ and $p$ for function space
Experiments with the JMTH dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>AP₀</th>
<th>AP₁</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
<th>Acc</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>balAPinc</td>
<td>0.57</td>
<td>0.56</td>
<td>0.573</td>
<td>0.573</td>
<td>0.573</td>
<td>57.3</td>
<td>53.8–60.7</td>
</tr>
<tr>
<td>ConVecs</td>
<td>0.76</td>
<td>0.77</td>
<td>0.703</td>
<td>0.702</td>
<td>0.702</td>
<td>70.2</td>
<td>66.9–73.3</td>
</tr>
<tr>
<td>SimDiffs</td>
<td>0.80</td>
<td>0.79</td>
<td>0.724</td>
<td>0.724</td>
<td>0.724</td>
<td>72.4</td>
<td>69.1–75.4</td>
</tr>
</tbody>
</table>

- balAPinc significantly below ConVecs and SimDiffs
- difference between ConVecs and SimDiffs not significant
Experiments with the JMTH dataset

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>AP$_0$</th>
<th>AP$_1$</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
<th>Acc</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.80</td>
<td>0.79</td>
<td>0.724</td>
<td>0.724</td>
<td>0.724</td>
<td>72.4</td>
<td>69.1–75.4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.76</td>
<td>0.75</td>
<td>0.680</td>
<td>0.680</td>
<td>0.680</td>
<td>68.0</td>
<td>64.6–71.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.79</td>
<td>0.79</td>
<td>0.717</td>
<td>0.716</td>
<td>0.716</td>
<td>71.6</td>
<td>68.3–74.7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.71</td>
<td>0.69</td>
<td>0.663</td>
<td>0.663</td>
<td>0.663</td>
<td>66.3</td>
<td>62.9–69.6</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>0.72</td>
<td>0.684</td>
<td>0.684</td>
<td>0.684</td>
<td>68.4</td>
<td>65.0–71.6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.76</td>
<td>0.74</td>
<td>0.690</td>
<td>0.690</td>
<td>0.690</td>
<td>69.0</td>
<td>65.7–72.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.75</td>
<td>0.73</td>
<td>0.701</td>
<td>0.701</td>
<td>0.701</td>
<td>70.1</td>
<td>66.8–73.2</td>
</tr>
</tbody>
</table>

- experiments with subsets of SimDiffs features
- all subsets contribute to performance
Experiments with the KDSZ dataset

- experimented with four ways of splitting the dataset
- (1) standard
  - standard 10-fold cross-validation
- (2) clustered
  - 10-fold CV but cluster pairs shared words in same fold
- (3) balanced
  - 10-fold CV but clustered and balanced classes
- (4) different
  - train on JMTH and test on KDSZ

- from easy to hard:
  - standard, clustered, balanced, different
Experiments with the KDSZ dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Evaluation</th>
<th>AP₀</th>
<th>AP₁</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
<th>Acc</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>balAPinc</td>
<td>standard</td>
<td>0.79</td>
<td>0.37</td>
<td>0.645</td>
<td>0.645</td>
<td>0.645</td>
<td>64.5</td>
<td>63.0–66.0</td>
</tr>
<tr>
<td></td>
<td>clustered</td>
<td>0.79</td>
<td>0.37</td>
<td>0.644</td>
<td>0.643</td>
<td>0.644</td>
<td>64.3</td>
<td>62.8–65.8</td>
</tr>
<tr>
<td></td>
<td>balanced</td>
<td>0.60</td>
<td>0.59</td>
<td>0.583</td>
<td>0.583</td>
<td>0.583</td>
<td>58.3</td>
<td>56.2–60.4</td>
</tr>
<tr>
<td></td>
<td>different</td>
<td>0.61</td>
<td>0.60</td>
<td>0.582</td>
<td>0.582</td>
<td>0.582</td>
<td>58.2</td>
<td>56.1–60.3</td>
</tr>
<tr>
<td>ConVecs</td>
<td>standard</td>
<td>0.87</td>
<td>0.56</td>
<td>0.731</td>
<td>0.747</td>
<td>0.735</td>
<td>74.7</td>
<td>73.3–76.1</td>
</tr>
<tr>
<td></td>
<td>clustered</td>
<td>0.78</td>
<td>0.36</td>
<td>0.636</td>
<td>0.690</td>
<td>0.645</td>
<td>69.0</td>
<td>67.5–70.5</td>
</tr>
<tr>
<td></td>
<td>balanced</td>
<td>0.60</td>
<td>0.59</td>
<td>0.567</td>
<td>0.554</td>
<td>0.531</td>
<td>55.4</td>
<td>53.3–57.5</td>
</tr>
<tr>
<td></td>
<td>different</td>
<td>0.57</td>
<td>0.62</td>
<td>0.569</td>
<td>0.561</td>
<td>0.547</td>
<td>56.1</td>
<td>54.0–58.2</td>
</tr>
<tr>
<td>SimDiffs</td>
<td>standard</td>
<td>0.88</td>
<td>0.60</td>
<td>0.749</td>
<td>0.757</td>
<td>0.752</td>
<td>75.7</td>
<td>74.3–77.0</td>
</tr>
<tr>
<td></td>
<td>clustered</td>
<td>0.80</td>
<td>0.40</td>
<td>0.664</td>
<td>0.684</td>
<td>0.671</td>
<td>68.4</td>
<td>66.9–69.9</td>
</tr>
<tr>
<td></td>
<td>balanced</td>
<td>0.63</td>
<td>0.64</td>
<td>0.596</td>
<td>0.592</td>
<td>0.588</td>
<td>59.2</td>
<td>57.1–61.3</td>
</tr>
<tr>
<td></td>
<td>different</td>
<td>0.58</td>
<td>0.61</td>
<td>0.581</td>
<td>0.574</td>
<td>0.564</td>
<td>57.4</td>
<td>55.3–59.5</td>
</tr>
</tbody>
</table>

- no significant differences on different evaluation
Experiments with the BBDS dataset

- experimented with three ways of splitting the dataset
  - (1) standard
    - standard 10-fold cross-validation
  - (2) clustered
    - 10-fold CV but cluster pairs shared words in same fold
  - (3) different
    - train on JMTH and test on KDSZ
  - no balanced split, because dataset is already balanced
Experiments with the BBDS dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Evaluation</th>
<th>AP₀</th>
<th>AP₁</th>
<th>Pre</th>
<th>Rec</th>
<th>F</th>
<th>Acc</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>balAPinc</td>
<td>standard</td>
<td>0.79</td>
<td>0.73</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
<td>72.2</td>
<td>70.5–73.8</td>
</tr>
<tr>
<td></td>
<td>clustered</td>
<td>0.79</td>
<td>0.73</td>
<td>0.722</td>
<td>0.722</td>
<td>0.722</td>
<td>72.2</td>
<td>70.5–73.8</td>
</tr>
<tr>
<td></td>
<td>different</td>
<td>0.79</td>
<td>0.73</td>
<td>0.701</td>
<td>0.687</td>
<td>0.682</td>
<td>68.7</td>
<td>67.0–70.4</td>
</tr>
<tr>
<td>ConVecs</td>
<td>standard</td>
<td>0.95</td>
<td>0.95</td>
<td>0.876</td>
<td>0.876</td>
<td>0.876</td>
<td>87.6</td>
<td>86.3–88.8</td>
</tr>
<tr>
<td></td>
<td>clustered</td>
<td>0.92</td>
<td>0.91</td>
<td>0.829</td>
<td>0.821</td>
<td>0.819</td>
<td>82.1</td>
<td>80.6–83.5</td>
</tr>
<tr>
<td></td>
<td>different</td>
<td>0.72</td>
<td>0.71</td>
<td>0.652</td>
<td>0.651</td>
<td>0.650</td>
<td>65.1</td>
<td>63.3–66.9</td>
</tr>
<tr>
<td>SimDiffs</td>
<td>standard</td>
<td>0.97</td>
<td>0.97</td>
<td>0.913</td>
<td>0.913</td>
<td>0.913</td>
<td>91.3</td>
<td>90.2–92.3</td>
</tr>
<tr>
<td></td>
<td>clustered</td>
<td>0.96</td>
<td>0.96</td>
<td>0.883</td>
<td>0.881</td>
<td>0.881</td>
<td>88.1</td>
<td>86.8–89.3</td>
</tr>
<tr>
<td></td>
<td>different</td>
<td>0.84</td>
<td>0.82</td>
<td>0.751</td>
<td>0.745</td>
<td>0.743</td>
<td>74.5</td>
<td>72.8–76.1</td>
</tr>
</tbody>
</table>

- all significant differences on *different* evaluation
# Experiments – Summary

- **bold** = significantly worse than SimDiffs
- no cases significantly better than SimDiffs

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>JMTTH Accuracy</th>
<th>KDSZ Accuracy</th>
<th>BBDS Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>balAPinc</td>
<td>57.3</td>
<td>58.2</td>
<td>68.7</td>
</tr>
<tr>
<td>ConVecs</td>
<td>70.2</td>
<td>56.1</td>
<td>65.1</td>
</tr>
<tr>
<td>SimDiffs</td>
<td>72.4</td>
<td>57.4</td>
<td>74.5</td>
</tr>
</tbody>
</table>
Discussion
Discussion of results

• results support the similarity differences hypothesis
  • second-order features useful for lexical entailment
• manually designing an asymmetric similarity measure is a difficult task
  • recall the complex equations for balAPinc
• manually designing feature vectors is easier and seems to work better
Limitations
Limitations and Future Work

• evaluation methodology
  • here: direct evaluation; future work: evaluate RLE module as component in larger RTE system

• variety of hypotheses
  • past work: mainly context inclusion hypothesis; here: context combination hypothesis, similarity differences hypothesis; future work: more hypotheses, combination of hypotheses

• semantic relations and lexical entailment
  • past work: no connection with work in semantic relations; here: strong connection with semantic relations; future: possible counter-examples for connection with semantic relations?
Conclusion
Conclusion

- SimDiffs better than balAPinc and ConVecs
- supervised learning better than manually building asymmetric similarity measures
- evidence for strong connection between lexical entailment and semantic relations