A Graph-based Approach for Contextual Text Normalization

EMNLP 2014, Doha, QATAR

Çağıl Uluşahin Sönmez & Arzucan Özdür
Boğaziçi University, Istanbul
Motivation

It's a beautiful night, looking for something fun to do, I think I want to be with my friends.
Text Normalization

<table>
<thead>
<tr>
<th>Imagine a world <strong>wer</strong> googling smt</th>
<th>Imagine a world <strong>where</strong> googling something</th>
</tr>
</thead>
<tbody>
<tr>
<td>u <strong>wer</strong> probly ekspektin sumthin</td>
<td>you <strong>were</strong> probably expecting something</td>
</tr>
</tbody>
</table>

- Two steps: **detection** and **normalization**
- Same non-standard token may be normalised differently depending on the **input context**
  - *e.g.* wer -> where or were
Related Work

<table>
<thead>
<tr>
<th>Method</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised Statistical Methods</td>
<td>Brill and Moore 2000, Toutanova et al. 2002,</td>
</tr>
<tr>
<td></td>
<td>Aw et al. 2006, Choudhury et al. 2007</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>Pennell and Liu, 2011</td>
</tr>
<tr>
<td>Normalization Lexicons</td>
<td>Han and Baldwin 2011, Han et al. 2012, Hassan</td>
</tr>
<tr>
<td></td>
<td>and Menezes 2013</td>
</tr>
</tbody>
</table>

- Supervised methods require huge amount of hand labeled data
- Lexicon based approaches always return same normalisation
- Most of them use no contextual information or only corpus based
- Cannot handle the evolving nature of social media text
Methodology
A Graph Based Approach for Contextual Text Normalization

• **Our proposed method:**

• An **Unsupervised** Text normalisation system based on Word Association Graph

• Uses a large **unlabelled** social media dataset

• Uses both corpus based contextual information & **input context** from input text.

• Also uses lexical similarity & grammatical features
Process Pipeline

- **Preprocessing**
  - Extract Candidates with Contextual Similarity Features
  - Extract Candidates from External Resources
  - Extract Candidates with Lexical Similarity Features

**Extract Candidates**

**Rank Candidates**
- Contextual Similarity Metrics
- External Score
- Lexical Similarity Metrics
Constructing the Word Association Graph
CWA-Graph is constructed using a large unlabeled corpus.

Models contextual information.

Nodes represent POS tagged words/tokens.

Both Out-of-Vocabulary (OOV) & In-Vocabulary (IV) words.

Edges encode relative positions of the POS tagged words.
Distance and Edge Weight

with Pronoun

a Determiner

distance: 0
weight: 25011

distance: 0
weight: 2918

distance: 0
weight: 305

distance: 1
weight: 198

distance: 1
weight: 322

distance: 2
weight: 89

beautiful Adjective

smile Noun

distance: 0
weight: 305
Normalization
Extracting Contextual Similarity Candidates
w Pronoun

a Determiner

beautiful Adjective

smile Noun

Distance: 0

neighbour\textsubscript{1}

neighbour\textsubscript{2}

eighbour\textsubscript{3}

out of vocabulary word\textsubscript{2}

candidates extracted from graph

new Adj

beautiful Adj

big Adj
A beautiful smile
w Pronoun

Distance: 1

broken Adj

Distance: 1

c1

c2

nice Adj

c2

c3

new Adj

Distance: 1

c3

c4

beautiful Adj

c4

c5

big Adj

Distance: 0

c5

c6

best Adj

Distance: 0

c6

c7

great Adj

Distance: 0

c7

smile Noun

Distance: 0

neighbour2

Distance: 0

neighbour3

beautiful Adj

Distance: 0

O2

neighbour1
w Pronoun

a Determiner

smile Noun

Distance: 0

Distance: 1

Distance: 0

neighbour₁

neighbour₂

neighbour₃

w

a

smile

Distance: 0

Distance: 0

Distance: 1
Process Pipeline

- with contextual features
- with lexical features
- from external resources
Choosing Among Candidate Normalizations

- Contextual similarity score is calculated using edgeWeight and frequency scores (conSimScore)

- Lexical similarity score (LexSimScore) is calculated using Longest Common Subsequence Ratio (LCSR) and Edit distance

- externalScore (in the slang dictionary or in the transliteration table or not)
Experiments & Results
Experiments

• Graph Generation: We extracted 1 GB of English tweets from Stanford’s 476 million Twitter Dataset

• The resulting graph contains over 100K nodes and 46M edges

• POS tagger: CMU Ark Tagger, which is a social media specific POS tagger achieving an accuracy of 95% over social media text.
Test Datasets

• First Dataset: LexNorm1.1 (Han and Baldwin 2011)
  - 549 tweets with 1184 manually annotated ill-formed OOV tokens

• Second Dataset: Pennell Trigram Dataset (Pennell and Liu 2014)
  - 985 trigrams from 1925 sentences and 985 manually annotated ill-formed OOV tokens
  - SMS-like Corpus: collected using only messages sent via SMS
# Results on LexNorm1.1

## System Components

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexSimScore</td>
<td>28.28</td>
<td>28.20</td>
<td>28.24</td>
</tr>
<tr>
<td>externalScore</td>
<td>64.69</td>
<td>64.52</td>
<td>64.60</td>
</tr>
<tr>
<td>lexSimScore + externalScore</td>
<td>77.22</td>
<td>77.02</td>
<td>77.12</td>
</tr>
</tbody>
</table>

## Methodology

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Han &amp; Baldwin, 2011</td>
<td>75.30</td>
<td>75.30</td>
<td>75.30</td>
</tr>
<tr>
<td>Liu et al., 2012</td>
<td>84.13</td>
<td>78.38</td>
<td>81.15</td>
</tr>
<tr>
<td>Yang et al., 2013</td>
<td>82.09</td>
<td>82.09</td>
<td>82.09</td>
</tr>
<tr>
<td>CWA-Graph, Sönmez &amp; Özgür</td>
<td>85.50</td>
<td>79.22</td>
<td>82.24</td>
</tr>
</tbody>
</table>
# Results on Trigram Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexSimScore</td>
<td>39.10</td>
<td>38.40</td>
<td>38.70</td>
</tr>
<tr>
<td>externalScore</td>
<td>44.20</td>
<td>43.30</td>
<td>43.80</td>
</tr>
<tr>
<td>lexSimScore + externalScore</td>
<td>65.50</td>
<td>65.20</td>
<td>65.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pennell and Liu, 2011</td>
<td>69.70</td>
<td>69.70</td>
<td>69.70</td>
</tr>
<tr>
<td>CWA-Graph, Sönmez &amp; Özgür</td>
<td>78.20</td>
<td>68.5</td>
<td>73.10</td>
</tr>
</tbody>
</table>
# Results without Pre-identification of OOVs

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Han et al., 2012</td>
<td>70.00</td>
<td>17.90</td>
<td>28.50</td>
</tr>
<tr>
<td>Hassan et al., 2013</td>
<td>85.37</td>
<td>56.40</td>
<td>69.93</td>
</tr>
<tr>
<td>CWA-Graph, Sönmez &amp; Özgür</td>
<td>85.50</td>
<td>79.22</td>
<td>82.24</td>
</tr>
</tbody>
</table>

Results obtained without assuming ill-formed words have been pre-identified.

- Baseline detection: tokens from graph with a frequency higher than 20 were filtered using GNU Aspell
Summary of Contributions

• **Unsupervised graph-based** text normalization approach

• Besides lexical and **grammatical** features, utilizes both corpus based and **input based contextual** features of social text

• **State-of-the-Art** Precision and F-scores
Future Work

• Doing Better OOV Detection
• Experimenting on Different Graph Sizes
• Performing Normalization for Other Languages
“Thank You”

– Çağıl Uluşahin Sönmez