A statistical model for morphology inspired by the Amis Language

Isabelle Bril, Achraf Lassoued, Michel de Rougemont
Lacito-CNRS, University Paris II, University Paris II & IRIF-CNRS
Plan

1. Morphology of the Amis language
   Vectors for words: Word2Vec
2. Statistical model of Morphology
   Factorization of vectors
3. Vectors for sentences
   Content vector
4. Best derivation trees in a Grammar
1. Morphology of natural languages

English: pre-existing pre-conceived pre-conception

Amis: mi-padang k-u taw ‘people help’ Actor voice
      AV-help NOM-ART people
padang-i k-aku ‘help me!’ Imperative Locative Voice
      help-IMP.LV NOM-1sg (lit. let me be helped)
ni-padang-en NMZ.PFV-help-PASS
pa-pi-padang ‘make s.o. help’
pa-pi-padang-en ‘make s.o. be helped’
Vectors for words: \( v_i \cdot v_j \approx M(i, j) \)

1.1 Basic statistics for words: PCA (Principal Components Analysis), for the Correlation matrix

- SVD (Singular Value) Decomposition
- Learning techniques reduce the dimension: from \(10^4\) to 200 for Word2Vec

1.2 Statistics for sentences:

- Stanford NLP group:
- Content vector based on the Morphology
Austronesian Languages

Gray et al., Nature 2009
Formosan Languages

- 14 different languages

- **Amis** has 4 variations from North to South.

Paul LI, Academia Sinica
2. Statistical model: Prefix distribution
(Prefix;Suffix) distribution of the root "banaq"

- (ma;)
- (ka;)
- (ni-ka;an)
Second Moments: Vector representation of prefixes

mi- padang t-u suwal n-ira tatakulaq.

.......

<table>
<thead>
<tr>
<th></th>
<th>k</th>
<th>ka</th>
<th>n</th>
<th>ni</th>
<th>mi</th>
<th>pa</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ka</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>n</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>ni</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>mi</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pa</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>t</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

PCA: Projection on 2 largest eigenvalues
Morphology based vectors: factorization

mi-padang  padang-i

Similarly for:

ni-padang-en

pa-pi-padang-en

19/09/2017
Structured distances

Classical distances:

\[ \text{vi} \cdot \text{vj} \cong M(i, j) \quad \text{dist}(\text{vi}, \text{vj}) \]

Structured distances

\[ \text{vi} \cdot \text{vj} \cong M_{\text{prefix}}(i, j) + M_{\text{roots}}(i, j) + M_{\text{suffix}}(i, j) \]

\[ \text{dist}_{\text{prefix}}(\text{vi}, \text{vj}), \quad \text{dist}_{\text{roots}}(\text{vi}, \text{vj}), \quad \text{dist}_{\text{suffix}}(\text{vi}, \text{vj}), \]
### 3. Vector for sentences: Content Vector

<table>
<thead>
<tr>
<th>Property</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valence</td>
<td>{0,1,2,3}</td>
</tr>
<tr>
<td>Voice</td>
<td>{AV,UV,LV,INST.V}</td>
</tr>
<tr>
<td>Tense</td>
<td>{Present,Past,Fut}</td>
</tr>
<tr>
<td>Mood</td>
<td>{Ind, Imp,Hort,Subj}</td>
</tr>
<tr>
<td>Illocutionary Force</td>
<td>{Decl, Neg,Exclam}</td>
</tr>
<tr>
<td>Information Structure</td>
<td>{Topic, Cleft Focus}</td>
</tr>
</tbody>
</table>

Mi-padang t-u suwal n-ira tatakulaq.

<table>
<thead>
<tr>
<th>Valence</th>
<th>[0,0,1,0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice</td>
<td>[1,0,0,0]</td>
</tr>
<tr>
<td>Tense</td>
<td>[.9,.05,.05]</td>
</tr>
<tr>
<td>Mood</td>
<td>[1,0,0,0]</td>
</tr>
<tr>
<td>Illocutionary Force</td>
<td>[1,0,0]</td>
</tr>
<tr>
<td>Information Structure</td>
<td>[0.5, 0.5]</td>
</tr>
</tbody>
</table>

C is a Probabilistic vector

Example: dimension 6

Uniform distribution
The Content vector is computed online: \( C_{i+1} = F(C_i, w_i) \)

**Amis:**

- Prefix \textit{mi-} determines Actor Voice AV
- Suffix \textit{–en} determines passive Voice UV

Hence \( \textit{mi-padang t-u suwal n-ira tatakulaq} \)

\( (he) \text{ supported the word of the frog} \)
4. Tree decompositions | Content Vector

- Fix a Grammar
  \[ S \rightarrow VP + VP.KP \]
  \[ VP \rightarrow Voice.V.KP + Voice.V \]

Each sentence may have an exponential number of tree decompositions.

Which tree is the most likely?

\[
\text{mi-padang} \ t-u \ suwal \ n-ira \ tatakulaq
\]
\( (he) \text{ supported the word of the frog} \)
\[ AV > \text{ hence } \ KP \text{ depends on VP} \]
\( \text{(case phrase)} \)
Tree decompositions | Content Vector

(a): KP depends on VP,  Most likely tree
Definition of “Most likely “

1. Stochastic grammars

\[
S \rightarrow \text{VP} \frac{2}{3} + \text{VP.KP} \frac{1}{3}
\]

\(\Omega_1\): probabilistic space on the sentences.  Find the most likely tree is hard! (#P hard)

2. Probabilistic Content

Vector C is a probabilistic space \(\Omega_2\)

Find t such that \(\text{Prob}[t \text{ is a tree decomposition } / \text{ C,....}]\) is Maximum
Conclusion

1. Amis has a strong grammatical morphology
   All languages have some morphology

2. Statistical vectors can be factorized

3. Content Vector for a sentence

4. Most likely tree decompositions
   The Content Vector defines a probabilistic space
   Most likely tree: approximation algorithm
Vectors from a Correlation Matrix

Correlation of two random variables: matrix

Word correlation in phrases: A B C A

PSD matrix = $M M'$, i.e. vectors

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>