Automatic animacy classification
for Dutch

Jelke Bloem, Gosse Bouma
Noun animacy

› Sentience of the referent

- sister – participant – carpenter – dude – northener
- cat – angel – dragon – bacteria
- oak – robot – community – government
- fountain – second – observation – music – fog

› Animacy hierarchy
Animacy hierarchy

› Humans

› Other animates
  ORGANIZATIONS, ANIMALS, INTELLIGENT MACHINES, VEHICLES.

› Inanimates
  CONCRETE INANIMATE, NON-CONCRETE INANIMATE, PLACE, TIME

› HUMAN > NONHUMAN > NONANIMATE
Animacy and grammaticality

› The spoon **which** is on the table is mine.
› * The man **which** is sitting on the table is my friend.

› * The spoon **who** is on the table is mine.
› **The man** **who** is sitting on the table is my friend.

*who* refers to **ANIMATE**, **which** refers to **INANIMATE**

› Cut-off point
Animacy and sentence processing

› Dative alternation  
(Bresnan et al. 2007)
She gave a push to the car.  
(prepositional dative)
She gave the car a push.  
(double object)
She gave a toy to the child.  
(prepositional dative)
She gave the child a toy.  
(double object)

› For inanimate recipients, the double object construction is used more often
Automatic animacy classification goals

› Corpus annotation
› Use in language technology
  • e.g. Automatic translation:
    De man die op de tafel zat
die = that, which, who, those, these
    The man who sat on the table
  • Anaphora resolution (Orasan and Evans, 2007)
    The tree fell on the man. He survived.
  • Better parsing (Øvrelid, 2009)
Animacy in natural language processing

- Few animacy resources are available (Zaenen, 2004)
- Therefore, few tools make use of animacy

- A few animacy classifiers were made, none for Dutch

- Dutch resources:
  - Cornetto (lexical-semantic database)
  - Lassy Large (automatically annotated corpus), 1.5 billion words
Animacy classification task

› For any noun, decide whether it refers to a human, nonhuman animate or inanimate entity

› Classification features
  • World knowledge?
  • Morphology?
  • Context?
Animacy classification task

› World knowledge  
  • Lexical-semantic database (WordNet)

    poet -> writer -> communicator -> person
    wikipedian -> ????

› Morphology

    诗  人
    poetry person

    or: Case marking
Context features

› Animates prefer the agent role & subject position
› Inanimates prefer the patient role & object position
› Genitive case
  • das Haus meines Vaters
› Reflexive
  • The teacher hurt himself
Context features

› Lexical association features: Verbs

The doctor thought John was right.
The banana thought John was right.

› Adjectives

The lazy thief.
The lazy hurricane.
Animacy: Data

› Word lemmas and their animacy from Cornetto

› Verb-argument relations from Lassy Large corpus

<noun animacy="nonanimate">gevoel</noun>
<noun animacy="nonanimate">IJsselmeer</noun>
<noun animacy="nonanimate">noord</noun>
<noun animacy="nonanimate">paasei</noun>
<noun animacy="human">doctor</noun>
<noun animacy="human">Engelsman</noun>
<noun animacy="human">roker</noun>
<noun animacy="human">symfonieorkest</noun>
<noun animacy="nonhuman">fuchsia</noun>
<noun animacy="nonhuman">pony</noun>
<noun animacy="nonhuman">yeti</noun>

85#blijf|intransitive|su#gevoel
298#ontsta|intransitive|su#gevoel
1#schrijf|transitive|su#gevoel
8#rest|intransitive|su#gevoel
7#ontdek|transitive|su#Engelsman
4#ontwerp|transitive|su#Engelsman
3#overschat|transitive|su#Engelsman
Classification procedure

- K-nearest neighbor (TiMBL)
- Each noun is a feature vector
- Classify new instances based on most similar (nearest) noun in multidimensional feature space
Classification procedure

› K-nearest neighbor (TiMBL)
› Each noun is a feature vector
› Classify new instances based on most similar (nearest) noun in multidimensional feature space

› 4 of 5 neighbors are inanimate
  • The inanimate class is assigned
Feature values: Association strength

› Noun-verb association
  Subj(\textit{ben}): bestuurder 125
  Subj(\textit{rij}): bestuurder 12

› Which is more interesting?

› Pointwise Mutual Information, Fisher’s Exact Test

Feature-noun pairs that co-occur more often than would be expected by chance
## Association strength

“gevoel” *(feeling, inanimate)* subject relation strength (Fisher’s)

<table>
<thead>
<tr>
<th>Association strength</th>
<th>Word</th>
<th>Association strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000000000000000</td>
<td>ontsta</td>
<td>arise</td>
</tr>
<tr>
<td>0.000000000000830</td>
<td>heb</td>
<td>have</td>
</tr>
<tr>
<td>0.000000000002380</td>
<td>speel</td>
<td>play</td>
</tr>
<tr>
<td>0.000000000501125</td>
<td>ben</td>
<td>be</td>
</tr>
<tr>
<td>0.0000003404273</td>
<td>zeg</td>
<td>say</td>
</tr>
<tr>
<td>0.731409478841741</td>
<td>krijg</td>
<td>get</td>
</tr>
<tr>
<td>0.823487761949459</td>
<td>spreek</td>
<td>speak</td>
</tr>
<tr>
<td>0.853510038160385</td>
<td>neem</td>
<td>take</td>
</tr>
<tr>
<td>0.902189553992116</td>
<td>ken</td>
<td>know</td>
</tr>
<tr>
<td>1.000000000002866</td>
<td>schrijf</td>
<td>write</td>
</tr>
</tbody>
</table>
Classification procedure

› K-nearest neighbor (TiMBL)
› Feature values are association scores

› Evaluate by classifying unseen nouns according to these features
## Results: Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>80.92%</td>
</tr>
<tr>
<td>Object/subject ratio</td>
<td>81.09%</td>
</tr>
<tr>
<td>Verb subject relations</td>
<td>91.06%</td>
</tr>
<tr>
<td>Verb object relations</td>
<td>91.20%</td>
</tr>
<tr>
<td>Adjective relations</td>
<td>88.91%</td>
</tr>
<tr>
<td><strong>Subj+Obj+Adj</strong></td>
<td><strong>93.34%</strong></td>
</tr>
</tbody>
</table>

**Baseline:** Classify everything as the majority class

**Ten-fold cross validation accuracy scores**
Noun frequency

Classification low-frequency nouns is generally more difficult

<table>
<thead>
<tr>
<th>Frequency cutoff</th>
<th>Baseline</th>
<th>Accuracy</th>
<th>Number of nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0</td>
<td>76.68%</td>
<td>83.39%</td>
<td>30.950</td>
</tr>
<tr>
<td>&gt;1</td>
<td>78.16%</td>
<td>90.27%</td>
<td>16.454</td>
</tr>
<tr>
<td>&gt;10</td>
<td>80.92%</td>
<td>93.34%</td>
<td>12.168</td>
</tr>
<tr>
<td>&gt;100</td>
<td>84.00%</td>
<td>91.22%</td>
<td>6.276</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>88.99%</td>
<td>88.62%</td>
<td>1.671</td>
</tr>
</tbody>
</table>
# Results: Class confusion

- **Classification errors**

<table>
<thead>
<tr>
<th>Predicted -&gt;</th>
<th>Human</th>
<th>Nonhuman</th>
<th>Nonanimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>151</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Nonhuman</td>
<td>0</td>
<td>2</td>
<td>53</td>
</tr>
<tr>
<td>Nonanimate</td>
<td>1</td>
<td>3</td>
<td>982</td>
</tr>
</tbody>
</table>

NONHUMAN class is only chosen correctly twice!
Results: Two-way classification

› Human/Nonhuman

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>85.57%</td>
</tr>
<tr>
<td>All</td>
<td>98.03%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pred -&gt;</th>
<th>Human</th>
<th>NonH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>152</td>
<td>23</td>
</tr>
<tr>
<td>NonH</td>
<td>1</td>
<td>1040</td>
</tr>
</tbody>
</table>

› Animate/Inanimate

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>80.92%</td>
</tr>
<tr>
<td>All</td>
<td>92.52%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pred -&gt;</th>
<th>Anim.</th>
<th>Inanim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anim.</td>
<td>155</td>
<td>75</td>
</tr>
<tr>
<td>Inanim.</td>
<td>16</td>
<td>970</td>
</tr>
</tbody>
</table>
Discussion

› Can classify over 90% of Dutch nouns correctly
  • The Cornetto “nonhuman animate” class cannot be classified well
› Corpus creation/annotation
› Applications (parser, anaphora resolution)

› Token-based instead of lemma-based (DutchSemCor)?
› Reduce resource requirements
  • Incorporate morphology, seed set
References


Cornetto lexical-semantic database: [http://www2.let.vu.nl/oz/cltl/cornetto/](http://www2.let.vu.nl/oz/cltl/cornetto/)

Lassy Large corpus: [http://www.let.rug.nl/vannoord/Lassy/](http://www.let.rug.nl/vannoord/Lassy/)
Questions?
Thank you for your attention
## Classes

### Human
- Brabander
- Eerste-Kamerlid
- afstammeling
- begeleidingsteam
- drieling
- ex-burgemeester
- geallieerden
- haantje-de-voorste
- juf
- oermens
- racist
- tachtiger

### Nonhuman
- ANWB
- appelboom
- brandweer
- cycloop
- dienstensector
- embryo
- familie
- ijsbergsla
- maatjesharing
- microbe
- olifant
- snackbar
- vrouwenrechten

### Inanimate
- Groningen
- Koninginnendag
- appel
- belastingkantoor
- compassie
- friettent
- gebarentaal
- keel
- orkaan
- robot
- sneltrein
- terrorisme
- zeewier
Russian case marking

*pervogo* (*acc=*gen) *studenta* (*acc=*gen) first student
‘the first student’

*pervyj* (*acc=*nom) *zakon* (*acc=*nom) first law
‘the first law’

Fraser and Corbett (1995)
Dutch Wh-clefts

a. *Wat ik leuk vind, is die tafel(GEN=COMM,-ANIMATE)*
   what i like, is that table

b. *Wat ik leuk vind, is dat huis(GEN=NEUT,-ANIMATE)*
   what i like, is that house

c. *Wie ik leuk vind, is dat kind(GEN=NEUT,+ANIMATE)*
   who i like, is that child

d. *Wie ik leuk vind, is die vrouw(GEN=COMM,+ANIMATE)*
   who i like, is that woman

› Found no good counter-examples in corpus search
Dutch quantifier suffixes

(De Swart et al., 2008)

De studenten hebben beide*(-n) het boek gelezen.
the students have both the book read
‘The students have both read the book.’

De boeken werden beide(*-n) door de studenten gelezen.
the books were both by the students read
‘Both books were read by the students.’

› In written Dutch
Fisher’s Exact Test: Contingency table

- The Fisher’s exact test is calculated using tables
- Totals are fixed

The noun “gevoel” (*feeling*) as a subject of the verb “ontstaan” (*to start, to arise*)

<table>
<thead>
<tr>
<th></th>
<th>gevoel</th>
<th>¬gevoel</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontstaan</td>
<td>298</td>
<td>5927</td>
<td>6225</td>
</tr>
<tr>
<td>¬ontstaan</td>
<td>405</td>
<td>111952</td>
<td>112357</td>
</tr>
<tr>
<td>Column totals</td>
<td>703</td>
<td>117879</td>
<td>118582</td>
</tr>
</tbody>
</table>

\[ p < 0.00001 \]
Dependence and independence

• The p-value can go both ways: Association strength

The noun “gevoel” (feeling) as a subject of the verb “schrijven” (to write)

<table>
<thead>
<tr>
<th></th>
<th>gevoel</th>
<th>¬gevoel</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>schrijven</td>
<td>1</td>
<td>299</td>
<td>300</td>
</tr>
<tr>
<td>¬schrijven</td>
<td>702</td>
<td>117578</td>
<td>118282</td>
</tr>
<tr>
<td>Column totals</td>
<td>703</td>
<td>117879</td>
<td>118582</td>
</tr>
</tbody>
</table>

p > 0.99999
Association strength

- This p-value can be used as a measure of association strength.
- A low value indicates a strong association, a high value indicates none.
- Because the totals are fixed, you cannot compare p-values from samples of different sizes.
Fisher’s exact test Hypothesis

› H₀: The noun x and the verb y are independent in subject relations
› H₁: The noun x occurs as a subject of the verb y more often than would be expected by chance
Calculating the value

- The p-value expresses the total probability of the observed distribution (table) and all the more extreme ones

<table>
<thead>
<tr>
<th></th>
<th>gevoel</th>
<th>¬gevoel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontstaan</td>
<td>298</td>
<td>5927</td>
</tr>
<tr>
<td>¬ontstaan</td>
<td>405</td>
<td>111952</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>gevoel</th>
<th>¬gevoel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontstaan</td>
<td>299</td>
<td>5926</td>
</tr>
<tr>
<td>¬ontstaan</td>
<td>404</td>
<td>111951</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>gevoel</th>
<th>¬gevoel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontstaan</td>
<td>300</td>
<td>5925</td>
</tr>
<tr>
<td>¬ontstaan</td>
<td>403</td>
<td>111950</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>gevoel</th>
<th>¬gevoel</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontstaan</td>
<td>301</td>
<td>5924</td>
</tr>
<tr>
<td>¬ontstaan</td>
<td>402</td>
<td>111949</td>
</tr>
</tbody>
</table>
Calculating the value

\[ P(n) = \frac{6225! \times 112357! \times 703! \times 117879!}{298! \times 5927! \times 405! \times 111952! \times 118582!} \]

\[ P(n + 1) = \frac{6225! \times 112357! \times 703! \times 117879!}{299! \times 5926! \times 404! \times 111951! \times 118582!} \]

\[ \text{etc} \]

\[ p = P(n) + P(n + 1) + P(n + 2) + \ldots \]

- \( A \) and \( B \) are associated more strongly than would be expected by chance (\( \alpha = 0.001 \))
## Association measure evaluation

<table>
<thead>
<tr>
<th>Measure of association</th>
<th>Correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointwise Mutual Information</td>
<td>93.33%</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>91.37%</td>
</tr>
<tr>
<td>Frequency</td>
<td>90.96%</td>
</tr>
<tr>
<td>None (Baseline)</td>
<td>80.92%</td>
</tr>
</tbody>
</table>
Number of features
Wrapped Progressive Sampling (van den Bosch, 2004)

- TiMBL has many parameters:
  - Nr. of nearest neighbours
  - Feature vector distance measure
  - Neighbour weighting
  - Feature weighting
- Wrapped Progressive Sampling can automatically converge to the optimal parameters for the data set
Appendix references

