Text-to-Pictograph Translation and Vice Versa
For People with Intellectual Disabilities

Leen Sevens
Vincent Vandeghinste
Ineke Schuurman
Frank Van Eynde

February 24, 2017

Centre for Computational Linguistics
KU Leuven
Content

• Introduction
• Text2Picto translation
• Picto2Text translation
• Able to Include
• Conclusion
Introduction
Augmentative and Alternative Communication

- Assist people with **Intellectual Disabilities**
- Allow them to **communicate** or **use the Internet**
- Increase life quality by **reducing social isolation**
- **Picture-based communication systems** are a form of AAC technology
Picture-based communication

Three types:
- Universal pictographs
- Emoticons
- Pictographs for people with disabilities
Picture-based communication

• Between two and five million people in EU
• Need for picture-based communication interfaces that enable social contact for illiterate and pre-literature users
• Interfaces should be
  – Easy to use
  – Configurable
  – Flexible
WAI-NOT Platform

- Enabling internet access for people with mental disabilities
- Specific **applications** for these users
  - Chat and e-mail client with the Text2Picto engine
  - Easy-to-Read News
  - Text-to-Speech
  - Games
- [www.wai-not.be](http://www.wai-not.be)
Pictograph sets

**Sclera pictographs**
- hond (dog)
- ik (I)
- hond-eten-geven (dog-food-feed)
- afwas-proper-terugbrengen (dishes-clean-return)

**Beta pictographs**
- hond (dog)
- ik (I)
- fruit-persen (squeeze-fruit)
- gebraden-kip (roasted-chicken)
Text2Picto
Example message

Mijn kat eet vis.
(My cat eats fish.)
Original system (baseline)

- WAI-NOT’s original system
- **Without any language technology**
- If input word matches *file name* of pictograph, a pictograph is shown
- Due to homonymy, this can be **wrong**!
- **Low coverage**
  - No morphological variation
  - Coverage of 41.33% for Beta (possibly wrong translations)
  - Much worse for Sclera: unusable
- **BLEU score of 0.00% for Sclera**
- **BLEU score of 5.93% for Beta**
WAI-NOT email corpus

• About 70,000 email messages sent with WAI-NOT

• **Three types of messages**
  – Emails written by **literate people** (teachers, parents, ...)
    • Broad vocabulary
    • Hardest to translate
  – Short messages by **the intended users** (largest part)
    • No punctuation
    • No capitalization
    • Several spelling errors
  – **Noisy messages**
    • Random clicking on pictographs
WAI-NOT email corpus

• Development set
  – 186 WAI-NOT messages
  – Used to test the system
  – Used to tune the system parameters

• Evaluation set
  – 50 messages
  – Average length: 20 words
Running example

“Hij is genezen” (He has recovered)
System description
Shallow linguistic analysis

1. **Tokenization**
   – Splitting off all punctuation apart from hyphen/dash
2. **Basic spelling correction**
   – One deletion, one insertion, one substitution
3. **Part-of-Speech tagging**
   – HunPos tagger trained on manually corrected data (2 million words)
4. **Sentence detection**
   – Punctuation-based
5. **Separable verb detection (Dutch-specific)**
   – Verb + particle are sometimes written as separate words
6. **Lemmatization**
   – Look up token/tag combination in corpus
   – Else, apply regular expression rules
Example after shallow linguistic analysis

message 1

sentence 1

word 1
- token: Hij
- tag: VNW(pers, pron, nomin, vol, 3, ev, masc)
- lemma: hij

word 2
- token: is
- tag: WW(pv, t gw, ev)
- lemma: zijn

word 3
- token: genezen
- tag: WW(vd, vrij, zonder)
- lemma: genezen
System description

Direct Route:
- e-mail in text
- Shallow Linguistic Analysis
  - per word
- Dictionary Lookup
  - yes: Personal Pronoun Treatment
  - no: Dictionary Lookup

Semantic Route:
- Semantic Analysis
  - per synset
- Semantics to Pictos
  - Pictograph Set
- Choose optimal path
  - e-mail in pictos
Semantic analysis

• Detect words indicating **negative polarity**
• Look up Cornetto **synsets** for each word
  – Like WordNet for Dutch
  – Contains relations between synsets (groups of synonymous words)
  – Synsets are linked to several **lemmas**
  – We added some Flemish lemmas
Example after semantic analysis

message 1

sentence 1

word 1
- token: Hij
- tag: VNW(pers, pron, nomin, vol, 3, ev, masc)
- lemma: hij

word 2
- token: is
- tag: WW(pv, tgw, ev)
- lemma: zijn

word 3
- token: genezen
- tag: WW(3d, vrij, zonder)
- lemma: genezen

lexunit 1
- id: r.v-10263
- synset: d.v-9050
- pos: VERB_INTRANS

lexunit 2
- id: r.v-10264
- synset: d.v-1336
- pos: VERB_REFLEX

lexunit \( n \)
System description
Linking pictographs to synsets

- **Manually linked** to Cornetto synsets
  - 5710 Sclera pictographs
  - 2760 Beta pictographs
- Sclera pictographs often depict **complex concepts**, such as
  - Verb + object
  - Noun + noun
  - ...
Using the synset links

- **mens (human)**
  - Hyperonymy: man, meneer (man, sir)
  - Hyperonymy: vrouw, dame (woman, lady)
  - Hyponymy: buurvrouw (female neighbor)
  - Xpos-NearSynonym: vrouwelijk (female)

- **man, meneer (man, sir)**
  - Antonymy: **vrouw, dame (woman, lady)**

- **vrouw, dame (woman, lady)**
  - Xpos-NearSynonym: **vrouwelijk (female)**

- **buurvrouw (female neighbor)**
System description
The direct route

• Not all words can be analyzed with Cornetto (only nouns, verbs, adjectives and adverbs)

• **Personal pronouns** are very frequent in e-mail messages and are not included in Cornetto: explicit treatment

• A translation mechanism that uses a **dictionary**
  • Token / tag / lemma -> instant translation into pictograph
Example after pictograph linking

message 1

word 1

<table>
<thead>
<tr>
<th>token</th>
<th>tag</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hij</td>
<td>VNW(pers, pron, nom in, vol.3, ev, masc)</td>
<td>hij</td>
</tr>
</tbody>
</table>

word 2

<table>
<thead>
<tr>
<th>token</th>
<th>tag</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>is</td>
<td>WW(pv, tw, ev)</td>
<td>zijn</td>
</tr>
</tbody>
</table>

word 3

<table>
<thead>
<tr>
<th>token</th>
<th>tag</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>genezen</td>
<td>WW(vd, vni, zonder)</td>
<td>genezen</td>
</tr>
</tbody>
</table>

picto-single

sick

feel better

penalty 0

penalty -1

penalty 2

TRUE

penalty 4

head synset d_v_9516
System description
Finding the optimal path

- A* search algorithm
- Uses **parameters** (tuned beforehand using a local hill climber)
  - Maximum penalty threshold
  - Hyperonym penalty
  - XPos synonymy penalty
  - Antonymy penalty
  - Wrong number parameter
  - No number
  - Out-of-vocabulary parameter
  - Direct route advantage
Evaluation

- **50 messages** (83 sentences)
- We made *reference translations* in Beta and Sclera
- Translation with focus on how the content can best be expressed in pictographs
Automatic evaluation

• Progressively activating features of the system
• Evaluated using
  – **BLEU**: most used Machine Translation metric
  – **NIST**: similar to BLEU, but less credit to high-frequency non-informative n-grams
  – **WER**: word error rate
  – **PER**: position-independent word error rate
• Without/with (automatic) spelling correction
### Automatic evaluation results

<table>
<thead>
<tr>
<th>Condition</th>
<th>No spelling correction</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
<td>WER</td>
<td>PER</td>
<td>BLEU</td>
<td>NIST</td>
<td>WER</td>
<td>PER</td>
<td>BLEU</td>
<td>NIST</td>
<td>WER</td>
<td>PER</td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>Sclera</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.00</td>
<td>1.43</td>
<td>96.27</td>
<td>92.13</td>
<td>0.00</td>
<td>1.38</td>
<td>99.00</td>
<td>95.58</td>
<td>0.00</td>
<td>1.84</td>
<td>97.51</td>
<td>94.20</td>
<td>0.00</td>
<td>1.84</td>
</tr>
<tr>
<td>Lemmatis.</td>
<td>0.18∗</td>
<td>1.68†</td>
<td>94.48</td>
<td>89.36</td>
<td>0.91∗</td>
<td>1.73†</td>
<td>93.92</td>
<td>88.81</td>
<td>0.24∗</td>
<td>2.29†</td>
<td>92.27</td>
<td>87.02</td>
<td>1.17†</td>
<td>3.68†</td>
</tr>
<tr>
<td>Direct</td>
<td>10.74†</td>
<td>2.93†</td>
<td>75.41</td>
<td>69.20</td>
<td>11.57†</td>
<td>3.05†</td>
<td>74.59</td>
<td>68.09</td>
<td>14.17†</td>
<td>3.68†</td>
<td>71.96</td>
<td>65.88</td>
<td>16.55†</td>
<td>3.97†</td>
</tr>
<tr>
<td>Synonyms</td>
<td>12.02∗</td>
<td>3.32†</td>
<td>70.58</td>
<td>63.26</td>
<td>13.24∗</td>
<td>3.41†</td>
<td>70.03</td>
<td>62.43</td>
<td>16.55†</td>
<td>3.97†</td>
<td>67.54</td>
<td>60.50</td>
<td>16.12</td>
<td>3.96</td>
</tr>
<tr>
<td>Relations</td>
<td>11.44</td>
<td>3.29</td>
<td>72.24</td>
<td>64.50</td>
<td>12.75</td>
<td>3.42</td>
<td>71.41</td>
<td>63.26</td>
<td>16.12</td>
<td>3.96</td>
<td>68.78</td>
<td>61.33</td>
<td>16.12</td>
<td>3.96</td>
</tr>
<tr>
<td>Beta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.95</td>
<td>2.29</td>
<td>80.76</td>
<td>72.21</td>
<td>0.94</td>
<td>2.40</td>
<td>81.10</td>
<td>71.99</td>
<td>0.70</td>
<td>2.81</td>
<td>79.19</td>
<td>69.85</td>
<td>1.14†</td>
<td>3.53†</td>
</tr>
<tr>
<td>Lemmatis.</td>
<td>0.77</td>
<td>2.90†</td>
<td>77.05</td>
<td>66.93</td>
<td>0.85†</td>
<td>3.01†</td>
<td>76.94</td>
<td>66.14</td>
<td>1.14†</td>
<td>3.53†</td>
<td>74.92</td>
<td>63.78</td>
<td>1.98†</td>
<td>4.43†</td>
</tr>
<tr>
<td>Direct</td>
<td>11.96†</td>
<td>3.65†</td>
<td>66.59</td>
<td>57.48</td>
<td>12.72†</td>
<td>3.76†</td>
<td>66.14</td>
<td>56.47</td>
<td>23.01†</td>
<td>5.00†</td>
<td>52.42</td>
<td>43.31</td>
<td>25.91†</td>
<td>5.17†</td>
</tr>
<tr>
<td>Synonyms</td>
<td>16.57†</td>
<td>4.12†</td>
<td>56.24</td>
<td>46.91</td>
<td>18.70†</td>
<td>4.28†</td>
<td>55.12</td>
<td>46.01</td>
<td>23.01†</td>
<td>5.00†</td>
<td>52.42</td>
<td>43.31</td>
<td>25.91†</td>
<td>5.17†</td>
</tr>
<tr>
<td>Relations</td>
<td>18.56∗</td>
<td>4.22†</td>
<td>56.47</td>
<td>47.24</td>
<td>20.11∗</td>
<td>4.40†</td>
<td>55.46</td>
<td>46.01</td>
<td>25.91†</td>
<td>5.17†</td>
<td>51.29</td>
<td>42.07</td>
<td>25.91†</td>
<td>5.17†</td>
</tr>
</tbody>
</table>

*p < 0.05, †p < 0.01
Manual evaluation

- One judge
  - Remove untranslated non-content words to allow calculating recall
  - Judge for every translated word whether the pictograph is correct, to calculate precision
## Manual evaluation results

<table>
<thead>
<tr>
<th>Condition</th>
<th>Precision</th>
<th>With proper names</th>
<th>Without proper names</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>F-Score</td>
<td>Recall</td>
</tr>
<tr>
<td>Sclera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>77.60%</td>
<td>41.42%</td>
<td>54.01%</td>
</tr>
<tr>
<td>Text2Picto</td>
<td>89.24%</td>
<td>86.23%</td>
<td>87.71%</td>
</tr>
<tr>
<td>Rel. improv.</td>
<td>15.00%</td>
<td>108.19%</td>
<td>62.39%</td>
</tr>
<tr>
<td>Beta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>82.73%</td>
<td>62.23%</td>
<td>71.03%</td>
</tr>
<tr>
<td>Text2Picto</td>
<td>85.91%</td>
<td>89.45%</td>
<td>87.64%</td>
</tr>
<tr>
<td>Rel. improv.</td>
<td>3.84%</td>
<td>43.73%</td>
<td>23.38%</td>
</tr>
</tbody>
</table>
Improvement #1: syntactic simplification

Ik ben woensdag een lieve baby gaan bezoeken en ik heb hem een papsfles gegeven.
I am Wednesday a cute baby go visit and I have him a bottle given.
‘On Wednesday I went to visit a very cute baby and I gave him the bottle.’
Improvement #1: syntactic simplification

Ik ben woensdag
I am Wednesday
‘On Wednesday

papfles gegeven.
bottle given.

 tease.'
Improvement #1: syntactic simplification

- **Solution:** automated syntactic simplification for pictograph translation:
  - Split long input sentences into several shorter ones
  - Convert passive constructions into active constructions
  - Adhere to Subject-Verb-Object (SVO) order
  - Convert relative clauses into independent clauses
  - Etc.
Improvement #1: syntactic simplification

Ik ben woensdag een lieve baby gaan bezoeken en ik heb hem een papfles gegeven.
I am Wednesday a cute baby go visit and I have him a bottle given.
‘On Wednesday I went to visit a very cute baby and I gave him the bottle.’
Improvement #2: improved spelling correction

- **New method:**
  - Generate spelling variants (for both phonetic errors and typographic errors)
  - Fuzzy matching (Machine Translation) techniques for finding the best combination of spelling variants
Improvement #3: word sense disambiguation

• The original system did not yet select the appropriate sense of a word
• The most frequent sense was chosen, which resulted in incorrect translations
Delicious. ;-}
Improvement #3: word sense disambiguation

- We implemented a word sense disambiguation (WSD) tool (created by Vossen et al. (2010) within the DutchSemCor project)
- The WSD scores are now added as new features of the synsets in the Text2Picto engine
- A high WSD score biases the selection of the pictograph toward the winning sense
Picto2Text

Aa Bb Cc
1+1 = 2
Example message

Sequence of pictographs

Sclera

Beta

Text

Mijn kat eet vis.
(My cat eats fish.)
Input methods

- **Original version (WAI-NOT):**
  - Two-level static hierarchy
  - Too many unordered pictographs on lowest level
  - Categorial inconsistency

- **New approach:**
  - Three-level static hierarchy
  - Dynamic pictograph prediction
Three-level static hierarchy

- **Top-level categories:**
  - Topic detection applied on WAI-NOT corpus
  - Latent Dirichlet Allocation

- **Subcategories:**
  - Based on WordNet (Cornetto) relations

- **Lowest level:**
  - Ordered by frequency in WAI-NOT corpus
  - Overruled by logical ordering: numbers, months, days, pairs of antonyms
Dynamic pictograph prediction

• **N-gram based prediction**
  – WAI-NOT corpus was automatically translated into Beta / Sclera pictographs (280K pictographs)
  – Build language models using SRILM for Beta / Sclera
  – Present the most likely next pictograph based on the two previous pictographs

• **Word association based prediction**
  – Retrieve a list of semantically similar words with the DISCO tool and translate them into pictographs
The user inserts pictographs

The prediction tool suggests new pictographs

to bark, to eat, to be,...
Pictograph to Natural Language Generation

- **Difficulties**
  - Pictograph-for-word correspondence will almost never provide acceptable output
  - Pictograph languages often lack pictographs for function words
  - A single pictograph often encodes information corresponding to multiple words with multiple inflected word forms
THE PROCESS OF MACHINE TRANSLATION.
System architecture

- **Step 1.** Take the file names of the pictographs

  mijn.png  kat.png  eten.png  vis.png
Step 2. Find the synsets that are connected to these file names and retrieve all the lemma’s that are contained within these synsets.

- **mijn.png**
  - Mijn
  - English: my

- **kat.png**
  - Kat, poes, dakhaas, kattin, kater, ...
  - English: cat, kitten, kitty, ...

- **eten.png**
  - Eten, tafelen, maaltijden
  - English: eat, consume, ...

- **vis.png**
  - Vis
  - English: fish
• **Step 3.** For every lemma, generate its paradigm (reverse lemmatization)

- *mijn* → *English: my*
- *kat* → *English: cat, cats, kitten, kittens,...*
- *poes* → *English: eat, eats, ate, consumed,...*
- *vis* → *English: fish*
• **Step 4.** For every noun, create variants with/without article

English: the cat, a cat, cat, the cats,...
• **Step 5.** Find the **most likely combination** of all these words by using an A* algorithm, based on a **trigram language model** (trained on very large Dutch corpora with the SRILM toolkit)
  – Europarl v.3 corpus (Koehn, 2005), the CGN (Oostdijk and Broeder, 2002), the CLEF corpus (Peters and Braschler, 2001), the DGT corpus (Steinberger et al., 2012) and Wikipedia entries

**Output:**
Mijn kat eet vis.
(My cat eats fish.)
## Results

<table>
<thead>
<tr>
<th>Condition</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sclera</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.0175</td>
<td>1.5934</td>
<td>76.4535</td>
</tr>
<tr>
<td>Rev. lem.</td>
<td>0.0178</td>
<td>1.6852</td>
<td>76.8411</td>
</tr>
<tr>
<td>Direct</td>
<td>0.0420</td>
<td>2.2564</td>
<td>66.9574</td>
</tr>
<tr>
<td>Synsets</td>
<td>0.0535</td>
<td>2.5426</td>
<td>65.9884</td>
</tr>
<tr>
<td>Articles</td>
<td>0.0593</td>
<td>2.8001</td>
<td>67.4419</td>
</tr>
<tr>
<td><strong>Beta</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.0518</td>
<td>2.767</td>
<td>70.4457</td>
</tr>
<tr>
<td>Rev. lem.</td>
<td>0.0653</td>
<td>3.0553</td>
<td>70.3488</td>
</tr>
<tr>
<td>Direct</td>
<td>0.0814</td>
<td>3.3365</td>
<td>63.0814</td>
</tr>
<tr>
<td>Synsets</td>
<td>0.0682</td>
<td>3.1417</td>
<td>61.4341</td>
</tr>
<tr>
<td>Articles</td>
<td>0.0739</td>
<td>3.4418</td>
<td>63.1783</td>
</tr>
</tbody>
</table>
Conclusions about Picto2text

• Large improvement over baseline
• Ample room for further improvement
• **Future improvements** (to do):
  – Explore neural net-based models
  – Explore rule-based models
  – Explore hybrid models
  – Automated grammar correction
Able to Include
Project description

• European project: March 2014 – April 2017
• **Goal:** Improve the living conditions of people with Intellectual Disabilities (ID)
  – Natural Language Processing tools can bring independency to people with ID
  – Ex. Social media websites, email
• **Accessibility Layer** based on three key technologies:
  – **Tech 1:** Text and content simplifier (Simplext): *English & Spanish*
  – **Tech 2:** Text-to-speech functionalities: *Dutch, English & Spanish*
  – **Tech 3:** Text-to-Pictograph and Pictograph-to-Text translation: *Dutch, English & Spanish*
• The Accessibility Layer can be used in combination with a number of other applications, such as Facebook (on smartphones and tablets)
• It is context-aware and customizable
Extending the tool toward other languages

• “Language-independent” design of the Text2Picto and Picto2Text tools:
  – The systems can easily be extended toward other languages
  – We made English and Spanish versions of the tools (Russian and French: in progress)
  – Thanks to the use of WordNets: easy transfer of pictographs to words in other languages
  – Add language-specific linguistic resources
  – Deal with language-specific issues
User tests

• Our partners (Belgium, UK, Spain) are testing the tools in three user scenarios:
  – **Labor integration:** email client, word processor, pdf reader
  – **Mobility:** guidance systems for people with ID
  – **Leisure within information society:** Facebook, Twitter, WhatsApp,...
User tests
User tests

• The outcome of the pilots affects our approach:
  – Ex. The need for **decent pictograph input methods**
  – Ex. The users wanted to see the original input words written next to the pictographs in the Text2Picto tool: **PictoParallel**
  – Ex. Long pictograph translations turned out to be rather confusing for the users: **syntactic simplification methods**
  – Ex. We are constantly **adding new pictographs** (on request)
Conclusion
Up to now

- **Large improvements** over baseline
  - Tested in vitro
  - Partly tested in vivo
  - Make system usable in real life

- Expansion to **new languages**
  - If there is a WordNet
  - If there are basic linguistic resources available
  - Not too hard
Future and ongoing

- Continue in vivo testing
- Improve Picto2Text
- Think of new applications, new target groups (the elderly, immigrants,...)
Contact:
leen.sevens@kuleuven.be
http://ccl.kuleuven.be/~leen/

Picto demo & publications:
http://picto.ccl.kuleuven.be