Dutch Machine Translation and the Avenue Project

Simon Zwarts
Macquarie University
Carnegie Mellon University

August 24, 2005
Introduction

I am a PhD student at the Centre for Language Technology at Macquarie University, Sydney Australia.

The last 2 months I have been working on a Dutch to English MT system in the Avenue project.
Abstract

Prototyping a Dutch-English Machine Translation system, within the Avenue project in 2 months time.

There are very few resources available, no suitable bilingual lexicon, no morphology, no bilingual Grammars

Domain of the project: the European Parlement Proceedings

**Europarl Corpus Extract**

in his reply, instead of arguing his case on the facts, saw fit to make me the target of a volley of insults which even his welsh origins cannot fully excuse. having described me as 'ponderous and immature', he went on to describe my intervention as 'empty'.
Biggest Challenges in Dutch-English

- Dutch has cross dependencies
- Dutch is almost verb final
- Compound nouns in Dutch
- Dutch verb particles jumping around in the sentence
- English has a different way of negating (using auxiliary verbs)
- Tenses are slightly different
Biggest Challenges in Dutch-English

- Dutch has cross dependencies
- Dutch is almost verb final
- Compound nouns in Dutch
- Dutch verb particles jumping around in the sentence
- English has a different way of negating (using auxiliary verbs)
- Tenses are slightly different
Lexicon - Resources available

- Bilingual sentence aligned corpus, Europarl (about 28 million words each language)
- Wall Street Journal with Part of Speech (English)
- WOTAN corpus with Part of Speech (Dutch)
- Tool for statistical word alignment and probabilities: GIZA++
Because of time limitations we have to put some limitations:

- No stemming or morphology
- No additional information on words, like case or agreement
Lexicon - Different Lexicons

In order of creation/need:

- Lexicon - cross (automatic) (124166 entries)
- Top500 (manual) (724 entries)
- Random (manual) (380 entries)
- Morphology (semi-automatic) (1698)
- verbparticles (manual) (477 entries)
- Lexicon - last (automatic) (66609 entries)
Lexicon - Automatic derivation

Lexicon - cross:

We use the GIZA++ word probabilities to bootstrap

GIZA++ was trained in both directions and lists all possible translation even very unlikely ones

We apply a threshold in both directions (hence cross)

<table>
<thead>
<tr>
<th>het</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>0.561</td>
</tr>
<tr>
<td>it</td>
<td>0.211</td>
</tr>
<tr>
<td>we</td>
<td>0.003</td>
</tr>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>0.830</td>
</tr>
<tr>
<td>het</td>
<td>0.054</td>
</tr>
<tr>
<td>hoe</td>
<td>0.00057</td>
</tr>
</tbody>
</table>
Lexicon - Automatic derivation

Lexicon - cross:

Now we need to assign a POS to build a rule for it, this means:

- POS downscaling WSJ (± 35 tags)
- POS downscaling WOTAN-lite (± 175 tags)

We reduce this to the following tagset:

NOUN VERB DET PRON ADJ ADV PREP CD

Words are assumed to have the same Dutch and English POS and are added for each POS they have in common.
Top500

Browsing through the automated lexicon there are lots and lots of errors
(And it’s cynical/weird too)

Automatic Derived Lexical items

NOUN::NOUN |: ["menszijn"] → ["violence"]
(Dutch: (the essence of) being human)

NOUN::NOUN |: ["waarden"] → ["ladies"]
(Dutch: (ethical) values)

Solution: Rely on our friend Zipf and his theories
Manual translation of the top 500 (Dutch) word (occurence in text) Added POS: CONJ AUXV
Probably the most boring lexicon of all

Mainly to cover all lexical gaps in the development set

Development test set size: 150 sentences

Now I had on average one unknown word a sentence

Possessive, Non-possessive pronoun fix (as far as not cover in the top500)

Contains all other incidental lexical items I’d liked to fix
Some *incidental* morphology only.
Most morphology is picked up by GIZA++ and is handled by just having separate lexical entries, however when morphology needs more than one word it fails. Comparative: Every ADJ in Dutch -er and having a GIZA++ link to *more* gets *more* before the word Superlative: idem, with Dutch -est and English *most* ("nijpendste" -> "serious") → ("nijpendste" -> "most serious")

ADJ → ADV: Dutch suffix -sch will result in an English +ly
ADJ: ("typisch") -> "typical") → ADV: ("typisch") -> ("typically")

There is much more morphology but no time to address these issues
### Verb particles

Some examples of nice Dutch verb particles:

<table>
<thead>
<tr>
<th>Dutch Verb Particles</th>
<th>English Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>door-gaan</td>
<td>through go</td>
</tr>
<tr>
<td>in-zien</td>
<td>in see</td>
</tr>
<tr>
<td>over-nemen</td>
<td>over take</td>
</tr>
<tr>
<td>bijdrage-leveren</td>
<td>contribution deliver</td>
</tr>
<tr>
<td>af-stemmen</td>
<td>off vote</td>
</tr>
<tr>
<td>aan-stippen</td>
<td>on “making dots”</td>
</tr>
<tr>
<td>eens-zijn</td>
<td>once be</td>
</tr>
<tr>
<td>voor-zitten</td>
<td>before sit</td>
</tr>
<tr>
<td></td>
<td>continue</td>
</tr>
<tr>
<td></td>
<td>realise</td>
</tr>
<tr>
<td></td>
<td>adopt</td>
</tr>
<tr>
<td></td>
<td>contribute</td>
</tr>
<tr>
<td></td>
<td>coordinate</td>
</tr>
<tr>
<td></td>
<td>note/emphasize</td>
</tr>
<tr>
<td></td>
<td>agree</td>
</tr>
<tr>
<td></td>
<td>chair (a meeting)</td>
</tr>
</tbody>
</table>

- Ik zit de vergadering van de europese unie van vandaag voor
- I sit the meeting of the european union of today before
- I chair the meeting of the european union today
Verb particles

The lexicon part of the solution is to translate the verb particles to the empty string and to remember on the verbs we are still waiting for a particle

VERB::VERB |:  ["gaan"] -> ["continue"]
(
   (x1::y1)
   ((x0 verbparticle)=door)
)

VERBPART::VERBPART |:  ["door"] -> ["""]
(
   (x1::y1)
   ((x0 verbparticle)=door)
)
There are no more lexical gaps in the development set but what about the final test? Something is better than nothing, even if it’s very well possible that is not correct.

Walk again through the word probability file and select the most likely translation for every word which is not yet in on of the lexicon.

Assigned most likely Dutch POS
Grammar: verb positions

Verb final behaviour:
<text> <verb> <text> <verb> <verb> ... 
*ik zou dat morgen graag gaan willen doen
*I would that tomorrow gladly go want do
*I would like to do that tomorrow

And in clauses:
...dat <text> <verb> <verb> ...
*hij zei dat ik dat morgen moet gaan doen
*he told me that I that tomorrow must go do
*he told me I have to do that tomorrow
Grammar rules

- default (di-/in-)transitive rules
- special category for the clauses where Dutch is verb final
- Because of long dependencies we want to parse NP as long as possible, rules which combined NP which comma’s considerably slows down the transfer engine
- special rules for building negations (anchored on lexical items as “not” etc)
- verb particles ... see next slide
The discussed verb particles and now the conclusion: the grammatical part

**VP::VP : [VERB NP] -> [VERB NP]**

```
(x1::y1)
(x2::y2)
((x1 verbparticle) = *UNDEFINED*)
(x0=x1)
```

**VP::VP : [VERB NP VERBPART] -> [VERB NP]**

```
(x1::y1)
(x2::y2)
((x1 verbparticle) =c (x3 verbparticle))
(x0=x1)
```
Grammar Stats

94 Rules Total

63 Rules without Lexical items

31 Rules with Lexical items, mainly to fix frequent occurring political proverbial phrases
## Translation of the Development set

<table>
<thead>
<tr>
<th>#</th>
<th>Translation</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>in the long term this leads almost certain to competitive battle and conflict of interests between the European Union and the Council of Europe.</td>
<td>op de lange termijn leidt dit vrijwel zeker tot concurrentiestrijd en belangenconflicten tussen de Europese Unie en de Raad van Europa</td>
<td>in the long term, this is almost bound to lead to a competitive battle and conflict of interests between the European Union and the Council of Europe.</td>
</tr>
<tr>
<td>21</td>
<td>across this three benefits is a triple responsibility</td>
<td>tegenover deze drie voordelen staat een drievoudige verantwoordelijkheid</td>
<td>the counterpart of this threefold benefit is a threefold responsibility</td>
</tr>
<tr>
<td>82</td>
<td>know you, Europe is barely a taking off aircraft: if it flies faster to too can raise, if it plunge to below. you have, Mr Prodi, the biggest responsibility.</td>
<td>weet u, Europa is net een opstijgend vliegtuig: of het vliegt sneller om te kunnen stijgen, of het sort naar beneden. u hebt, mijnheer Prodi, de grootste verantwoordelijkheid</td>
<td>the truth is that Europe resembles an aeroplane taking off: either it accelerates and keeps rising upwards or it falls to the ground. and so, president, you who bear the greater responsibility</td>
</tr>
</tbody>
</table>
Some Evaluations metrics on the development set

<table>
<thead>
<tr>
<th>Metric</th>
<th>Precision</th>
<th>Recall</th>
<th>Penalty</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.2289</td>
<td>—</td>
<td>0.9745</td>
<td>0.1640</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.2040</td>
<td>0.2017</td>
<td>—</td>
<td>0.2028</td>
</tr>
<tr>
<td>Meteor</td>
<td>0.5150</td>
<td>0.5132</td>
<td>0.1335</td>
<td>0.4447</td>
</tr>
</tbody>
</table>
Questions?