COLIBRI: Constructions as Linguistic Bridges

Maarten van Gompel

December 2011

Background information

- *2008-2009*: Master student Human Aspects of Information Technology at Tilburg University
- *2009-2011*: Scientific programmer at Tilburg University
- *2011–*: PhD researcher at RU since September 2011

My field

- Language Technology
  - Implicit Linguistics
  - Machine Learning
  - Machine Translation

Translation is hard

Example

Automated translation is difficult!

Nederlands: De PVV wil fors korten op ontwikkelingssamenwerking. De peiling van De Hond geeft aan dat slechts 4 procent van zijn achterban dat absoluut niet wil. Bijna een op de vijf CDA-stemmers is daar echt niet voor te vinden, terwijl voor het CDA in de Tweede Kamer verlagen van ontwikkelingshulp moeilijk ligt. (`bron: nu.nl`)

Google Translate: The PVV will considerably shorten the development. The poll of Dog indicates that only 4 percent of his supporters that absolutely does not want. Nearly one in five voters CDA is really not to be found, while the Christian Democrats in the House reduction of development is difficult.

Translation is hard

Naive word-by-word? No

I	see		a	poor	man
	Veo	a	un	hombre	pobre

Idioms

- It's a piece of cake!
- Het is een stukje taart!
- Het is een peulenschil!
- It's a pea shell!

Translation is hard

Ambiguity

- Mijn geld staat op de **bank**
- My money is in the **bank**

- Mijn geld ligt in een doosje onder de **bank**
- My money is in a box under the **sofa**

- My boat is on the **bank** of the river
- Mijn boot ligt aan de **oever** van de rivier

Approaches

1. **Rule-based:** explicit linguistic knowledge
2. **Data-driven:** implicit linguistics
    - Statistical models
    - Memory-based machine learning

Two important aspects of translation

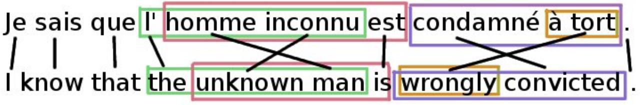
1. Faithful conservation of meaning
2. Fluent natural style

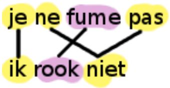
These can be modelled

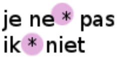
1. Faithful conservation of meaning: $argmax_T P(T|S)$
2. Fluent natural style: $argmax_T P(T)$
3. $besttranslation_T = argmax_T P(T) \cdot P(T|S)$

What are the units of translation? → constructions

- Whole sentences at once? No
- Single words, word by word? No
- Words in context? Better
- Variable-length phrases in context? Even better
- Loosening constraints: “constructions” in context? Best?


Je sais que l' homme inconnu est condamné à tort .
I know that the unknown man is wrongly convicted .





What is a construction?

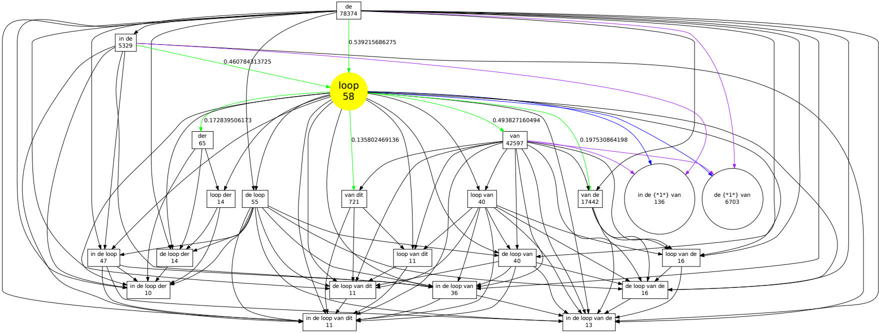
First stage of research:

The identification and extraction of "good" constructions from corpus data

We are looking for the intuitive "atom" in memory-based machine translation. A level above n-gram models and below syntactic level.

What is a construction?

- A *pattern* of words, possibly with "skips", which in some way forms an entity
- Constructions emerge from the data (parallel corpora) rather than linguistic theory:
    1. by frequency
    2. by subsumption
    3. by multilingual alignment


de
78374
0.539215686275
in de
5329
0.460784313725
loop
58
0.172839506173
0.493827160494
0.135802469136
0.197530864198
der
65
van
42597
loop der
14
de loop
55
van dit
721
loop van
40
van de
17442
in de {*1*} van
136
de {*1*} van
6703
in de loop
47
de loop der
14
loop van dit
11
de loop van
40
loop van de
16
in de loop der
10
de loop van dit
11
in de loop van
36
de loop van de
16
in de loop van dit
11
in de loop van de
13

Aligning constructions

Second stage of research:
Establishing aligned constructions, a mapping between constructions in two (or more) languages
How? *Measures of co-occurrence*

Example

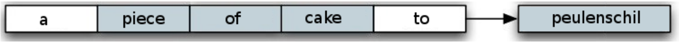
'Uraa al-kalba	I see the dog
'Uraa al-qitta	I see the cat
akala al-kalbu	The dog ate
Yaqtulu al-kalbu al-qitta	The dog kills the cat

Try now: *'uraa? al-kalbu? al-qitta? akala? yaqtulu? al?*

Machine Learning - Construction experts

Third stage: *Local translation step*
Once we have alignments, can we build automated classifiers that 'predict' the translation of constructions?
Hypothesis: Construction-experts, one classifier for each construction, increase translation accuracy.

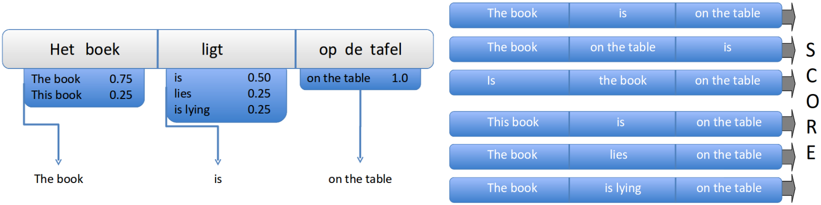
Predicted translation is made on the basis of to-be-determined features, such as context.



Decoding

Fourth stage: *Global translation step – Decoding*

- **search problem:** Recombine translated constructions into coherent translations in the target language.
- Translation hypothesis are scored using some kind of evaluation function

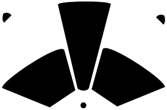


Het boek		ligt		op de tafel	
The book	0.75	is	0.50	on the table	1.0
This book	0.25	lies	0.25		
		is lying	0.25		

The book	is	on the table
The book	on the table	is
Is	the book	on the table
This book	is	on the table
The book	lies	on the table
The book	is lying	on the table

Evaluation

- Full implementation needed before we can evaluate properly
- Evaluation of Machine Translation is not trivial.
  - Comparison to human reference translations



Questions?