Fries in data-gestuurde automatische vertaling

Maarten van Gompel, Radboud University Nijmegen

Juni 2012

Automatisch vertalen Fries-Nederlands

Doelstelling: Het ontwikkelen van een automatisch vertaalsysteem voor Fries-Nederlands en Nederlands-Fries

Samenwerking

- Fryske Akademy
- Radboud Universiteit Nijmegen
  - Prof. dr. Antal van den Bosch – Hoogleraar example-based language modelling
  - Maarten van Gompel – Assistent in Opleiding "Constructions as Linguistic Bridges"
- TextInfo B.V.

Vertalen is moeilijk

Example

Automatisch vertalen is moeilijk

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.

Vertaling is moeilijk

Woord-voor-woord? Nee

I	see		a	poor	man
	Veo	a	un	hombre	pobre

Idiomen

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

Vertalen is moeilijk

Ambiguïteit

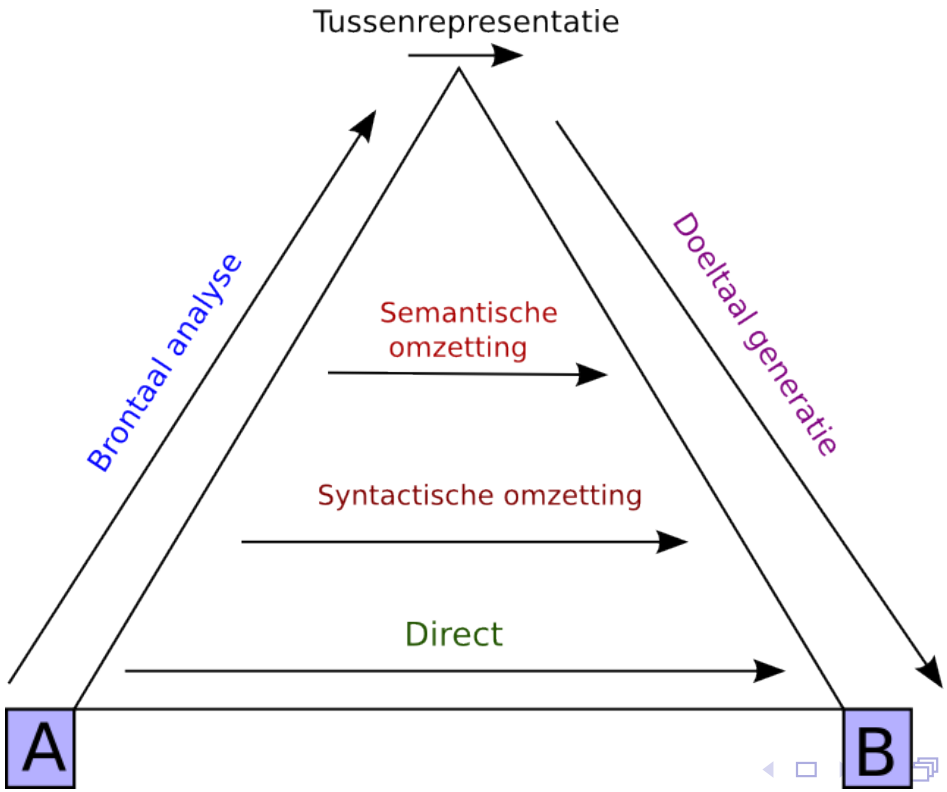
- Mijn geld staat op de **bank**
- My money is in the **bank**

- 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

Methoden

1. **Regel-gebaseerd:** expliciete linguistische kennis
2. **Data-gestuurd:** "impliciete linguistiek"
   - Statistische modellen
   - Machine learning



Data-gestuurde vertaling

Data gestuurde vertaling – automatisch leren op basis van een parallel corpus

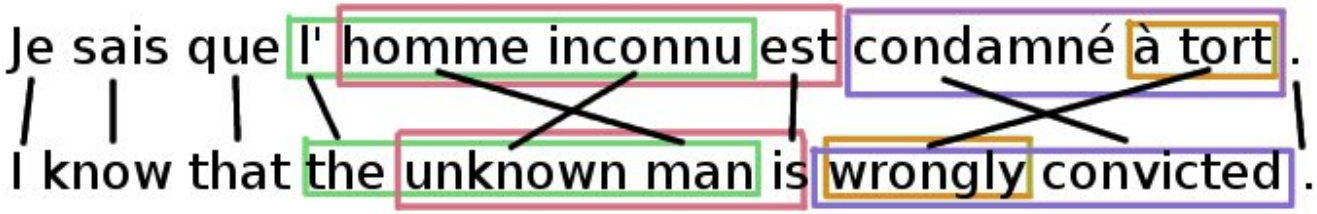
1. Fryske Akademy levert parallelle teksten aan – Fries/Nederlands
2. Wij vinden algoritmisch welke zinnen vertalingen zijn (momenteel 30.000 zinnen)
3. We extraheren paren van naar elkaar vertalende **frases**
4. Deze vormen het geleerde vertaalmodel

Uitdagingen bij Nederlands-Fries

- Is er genoeg data?
- Gevarieerd corpus?
- Dialectische variëteit?

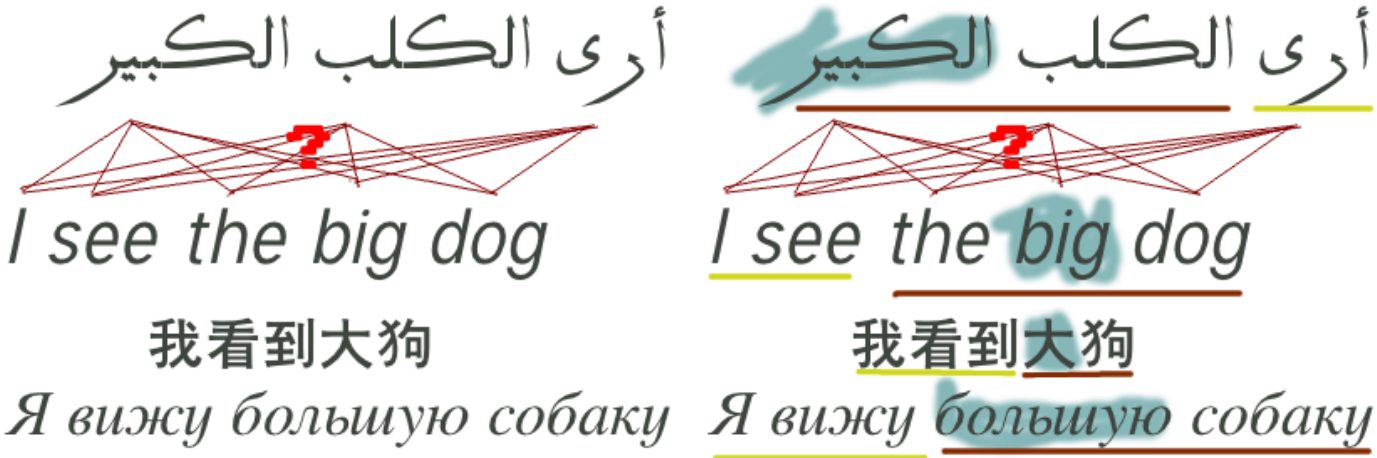
Wat zijn de eenheden van vertaling → frases

- Hele zinnen? Nee
- Losse woorden? Nee
- Woorden in context? Beter
- Frases van variabele lengte? Ja



Hoe kunnen vertalende frases gevonden worden?

Hoe? *Mate van samen voorkomen*



Example

'Uraa al-kalba al-kabira	Ik zie de grote hond
'Uraa al-qitta al-saghira	Ik zie de kleine kat
'Uraa al-qitta al-kabira	Ik zie de grote kat
akala al-rajul	De man at
Yuhabbu al-rajul al-qitta	De man houdt van de kat

Probeer!

'uraa?

al-kalba?

al-qitta al-saghira?

al-rajul?

yuhabbu?

Modellering –Twee belangrijke aspecten van vertaling

1. Behoud van betekenis
2. Natuurlijke stijl

Deze kunnen statistisch benaderd worden

1. Behoud van betekenis: $argmax_T P(T|S)$
2. Natuurlijke stijl: $argmax_T P(T)$
3. $vertaling_T = argmax_T P(T) \cdot P(T|S)$

Het vertaalsysteem

Gegeven een nieuw te vertalen zin en het geleerde model:

1. Zoek hierin alle voorkomens van frases uit ons model
2. Zet deze voorkomens om naar de vertalingen, zodanig dat:
3. ... behoud van betekenis maximaal is
4. ... natuurlijke stijl maximaal is
5. $(vertaling_T = argmax_T P(T) \cdot P(T|S))$

Software

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst
Moses: Open Source Toolkit for Statistical Machine Translation Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Evaluatie

Evaluatie is niet triviaal; automatische evaluatie door overlap met referentievertalingen.

Eerste resultaat

Eerste resultaten zijn bemoedigend

- Fries-Nederlands, BLEU score: 0.49
- Nederlands-Fries, BLEU score: 0.46

Voorbeeld-vertalingen:

- Wij waren met zijn vieren als kinderen , twee zusters en twee broers .
- Wy wiene mei ús fjouweren as bern , twa susters en twa broers .
- Mar it foel my daliks op dat er s fernuvere oanseach .
- Maar het viel me meteen op dat hij ons verbaasde aankeek .
- “ Hast him noait sjoen ? ” frege Linda .
- “ Heb je hem nooit gezien ? ” vroeg Linda .

Conclusie & Toekomst

Conclusies

- Automatisch vertalen is moeilijk
- Bemoedigende eerste resultaten!
- Beperkte trainingsdata levert al aardige vertalingen.

Toekomst

- Meer data $\rightarrow$ betere vertalingen
- Integratie in webapplicatie (TextInfo B.V)

Tot slot ...Lancering Oersetter.nl