AuCoPro

Project Presentation and Recent Developments

Presented at CTexT – Centre For Text Technology, North-West University, Potchefstroom, South-Africa

7 September 2012

Ben Verhoeven

ben.verhoeven@student.ua.ac.be

Introduction

- **Au**tomatic **Co**mpound **Pro**cessing
- Compound = two words that form a new word
  - e.g. *coffee bean*

1. Compound splitting
   - e.g. *coffee* + *bean*
2. Analysis of compound semantics
   - e.g. 'bean that is used to make coffee'

Project Funding

- Nederlandse Taalunie – Flanders (Belgium) & The Netherlands
  - Dutch Language Union
- Department of Arts and Culture – South Africa

- Period: 2012-2013
- Goal of fund: Research on Dutch and South African languages, spreading of NLP knowledge and initiating more international contacts

Project Leaders

- Dr. Menno Van Zaanen (Tilburg University, The Netherlands)
  - Compound splitting
- Prof. Dr. Walter Daelemans (University of Antwerp, Belgium)
  - Compound semantics
- Prof. Dr. Gerhard Van Huyssteen (North-West University, South-Africa)
  - Linguistics
  - Project coordination

Compound Semantics: Collaborators

- Walter Daelemans
  - Project leader – Dutch
- Gerhard Van Huyssteen
  - Project leader – Afrikaans
- Ben Verhoeven
  - Computational linguist
  - Interim project leader
  - Co-mentor of four Bachelor's mini-dissertations
    - Nadia Schultz: resource managing
    - Carli De Wet: Afrikaans semantics of N+N compounds
    - Joanie Liversage: semantics of X+N compounds
    - Benito Trollip: morphological classification of Afrikaans compounds
- Dirk Snyman
  - Computational linguist (mostly from January 2013 onwards)
- Martin Puttkammer & Martin Schlemmer
  - Help where needed

Why Semantic Analysis?

- Meaning is often not clear on its own (ambiguity)
- Implicit semantic relation between constituents
  - e.g. *donut seat*
    - 'donut-shaped seat'?
    - 'seat with a donut nearby'?
- Natural language understanding
  - Machine translation
    - Paraphrase may be needed
    - e.g. *Antwerp hostel* (Eng) -> *Auberge à Anvers* (Fr)
  - Information retrieval
  - Information extraction
  - Question answering

Research Method (1)

- Classification experiment of N+N compounds
- 11 classes of compounds that describe relation between constituents
- Of which 6 semantically specific
  - BE e.g. *woman doctor*
  - HAVE *car door*
  - IN *garden party*
  - ACTOR *student manifestation*
  - INST *hammer blow*
  - ABOUT *stamp collection*
- Annotation of 1500-2000 compounds

Research Method (2)

Information for classification decision

- Context from corpus for every constituent
  - Implicit semantic representation of a word
  - Dutch: Twente News Corpus (TwNC)

- "You shall know a word by the company it keeps" (Firth, J. R. 1957:11)

Firth, J.R. (1957) *Papers in Linguistics 1934–1951*. London: Oxford University Press.

	Leash	Walk	Run	Owner	Pet	Bark
Dog	3	5	2	5	3	2
Cat	0	3	3	2	3	0
Lion	0	3	2	0	1	0
Light	0	0	0	0	0	0
Bark	1	0	0	2	1	0
Car	0	0	1	3	0	0

(Baroni, 2008)

Research Method (3)

- Machine learning
  - Support vector machines (SVM)
  - Instance-based learning

- Principal components analysis (PCA)
  - Mathematical method of dimensionality reduction of information

Accomplishments

- Adaptation of annotation guidelines for Dutch and Afrikaans
- Development of annotated lists of N+N compounds in Dutch and Afrikaans (alpha release)
- Initial results on Dutch compounds
- Development of new compound list (N+N) for Dutch with sample sentences, yet to be annotated

Initial Results on Dutch Compounds

- IAA of 60.2% (K=0.60)
- Highest *F*-score = 49.0%
  - vs. baseline of 29.5% -> SIGNIFICANT

- Annotation improvements possible:
  - Better education of annotators
  - Optimisation of guidelines
  - Annotation in context (with sample sentences)

Future Research: Afrikaans

Same Experiment

- Preliminary IAA of 53%
- Context statistics to be calculated on Media24 corpus (news corpus similar to TwNC)
- Creation of new compound list (N+N) with sample sentences
  - Compounds from CKarma
  - Sample sentences from Taalkommissie corpus (TK)

Future Research

- Other kinds of compounds
  - V+N
  - A+N
  - P+N
    - POS annotation of CKarma
- New annotations with sample sentences and improved guidelines

Deliverables

Accomplished

- Annotation guidelines for Dutch and Afrikaans
- Master's dissertation
- Annotated lists of N+N compounds in Dutch and Afrikaans (alpha release)

Future

- Annotated compound lists (with sample sentences) for Dutch and Afrikaans (beta release)
- (Partially) POS tagged CKarma
- Four Bachelor's mini-dissertations
- Paper for PRASA: present initial results
- Talk at CLIN
- Paper at bigger conference