AuCoPro

Automatic Compound Processing

http://www.tinyurl.com/aucopro

Semantic Classification of Dutch and Afrikaans Noun-Noun Compounds

Ben Verhoeven[1], Walter Daelemans[1] & Gerhard van Huyssteen[2]

[1] CLiPS – CLG, University of Antwerp, Belgium
{Ben.Verhoeven;Walter.Daelemans}@uantwerpen.be

[2] CTexT, NWU, Potchefstroom, South Africa
Gerhard.VanHuyssteen@nwu.ac.za

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Presented at AfLaT 2013
Ghent, Belgium

@clipsua

06/12/2013

Introduction

- Productivity of a language to create new words
  - Obstacle for computational language understanding
- Meaning of compound 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’
    - ‘seat made of donuts’ ?

Related Research (1)

- Focus on
  - English
  - Noun-noun compounds

- Supervised machine learning problem
- Predefined inventory of classes of semantic relations between constituents of compound

Related Research (2) Classification

- Two kinds of classification schemes
  - Paraphrasing preposition
    - E.g. *autodeur* = deur VAN auto
  - Predicate-based classes
    - Class AGENT: ‘X is performed by Y’
      - E.g. *studentenprotest* = protest performed by students

Related Research (4)
Features

- Taxonomy-based methods
  - Semantic network similarity
  - Word’s location in hierarchy of terms
    - E.g. Hyponomy in WordNet
      - E.g. cola < frisdrank< drank < vloeistof
- Corpus-based methods

Related Research (5)
Features

- Taxonomy-based methods
- Corpus-based methods
  - Co-occurrence information of constituents in corpus
  - Distributional hypothesis (Harris)
    - Set of contexts in which a word occurs is an implicit representation of its semantics

Annotation (1)

- Semantic information on compounds needed for machine learning
- Explicit description by manual annotation
- Constraints on compound selection
  - Not in dictionary
    - Otherwise, gloss already present
    - Train classifier on systematics of newly produced compounds
  - Constituents in dictionary
    - Semantically relating of unknown words seems pointless

Annotation (2)
Scheme and Guidelines

- Adopted from Ó Séaghdha (2008), adapted for Afrikaans and Dutch
- 11 classes of compounds that describe relation between constituents
- Of which 6 semantically specific

- BE	e.g.	*zanger-muzikant*	*skrywer-boer*
- HAVE		*autodeur*	*blomsteel*
- IN		*tuinfeest*	*nagaktiwiteite*
- ACTOR		*studentenprotest*	*beerjagter*
- INST		*hamerslag*	*tapytborsel*
- ABOUT		*postzegelverzameling*	*kategismusvrae*

Annotation (3)
Process

Dutch

- Compound list from e-Lex
- 1802 noun-noun compounds

- Second annotator: 500
- IAA = 60.2 % (Kappa = 0.60)

Afrikaans

- 1500 noun-noun compounds manually selected from Ckarma

- 3 annotators
- IAA = 53.4% (Kappa = 0.53)

Experiment (1)

- Ó Séaghdha (2008) as inspiration

- Lexical similarity
  - Compounds are semantically similar when their respective constituents are semantically similar
  - E.g. *mieliesak* ‘corn bag’ and *graanblik* ‘can of grain’

Experiment (2)
Vector Creation

- Co-occurrence context for every compound constituent
  - For each instance of constituent, *n* surrounding words were held in memory
  - Size of context: 3 & 5 left and right
  - Relative frequencies of context words stored in vector

- Twente News Corpus (Dutch): 340 million words
- Taalkommisiekorpus (Afrikaans): 60 million words

Experiment (3)
Vector Creation

- Instance vectors are concatenation of constituent data
- Relative frequencies for the 1000 most frequent words per constituent (2000 per compound)
- Experiment only on compounds in semantically specific classes
  - BE, HAVE, ABOUT, IN, ACTOR, INST

Principal Component Analysis (PCA)

- Size of vectors: 2000 attributes
- Computationally expensive
- PCA mathematically reduces dimensionality while optimising variance in data
- Correlated attributes are fused into principal components (PCs)
- For now: restriction to 50 PCs

Baseline

- First research for these languages
- Majority baseline, thus:
  - For Dutch: 29.5% (428/1447 class IN)
  - For Afrikaans: 28.2% (407/1439 class ABOUT)

Initial Results

DUTCH	P	R	F
BOW 3	47.1	47.9	47.3
BOW 5	46.7	47.8	47.1
PCA 3	43.7	47.3	43.7
PCA 5	42.9	48.0	43.2
Baseline	29.5		

AFR	P	R	F
BOW 3	50.8	51.6	51.1
BOW 5	50.3	50.8	50.5
PCA 5	49.3	51.3	48.5
PCA 3	47.7	50.5	47.5
Baseline	28.2		

Results of SVM on Dutch and Afrikaans compound semantics, using 10-fold cross-validation

- BOW and PCA[50]
- Size of context: 3 & 5

Initial Discussion

- Both languages show significant improvement over majority baseline
- BOW seems to do better than PCA

- Better results for Afrikaans
  - Possibly due to annotated list being a combination of semantic annotations of 3 persons
  - Most agreed upon class for each compound
- Dutch: just one annotator

Per-class performance

Dutch BOW 3

Category	F-Score
IN	60.1
ABOUT	52.9
HAVE	36.3
INST	40.6
BE	17.0
ACTOR	42.9
Average	*47.3*

IN is best performing category

BE does significantly worse than others

Per-class performance

Dutch BOW 3

Category	F-Score	Distribution
IN	60.1	29.5 %
ABOUT	52.9	26.6 %
HAVE	36.3	16.1 %
INST	40.6	16.2 %
BE	17.0	7.3 %
ACTOR	42.9	4.3 %
Average	*47.3*	

Afrikaans BOW 3

Category	F-Score	Distribution
IN	51.8	20.8 %
ABOUT	61.3	28.2 %
HAVE	23.9	9.7 %
INST	13.6	7.5 %
BE	56.9	25.0 %
ACTOR	62.2	8.8 %
Average	*51.1*	

Classes with fewer instances seem harder to learn

Easily learnable class: ACTOR

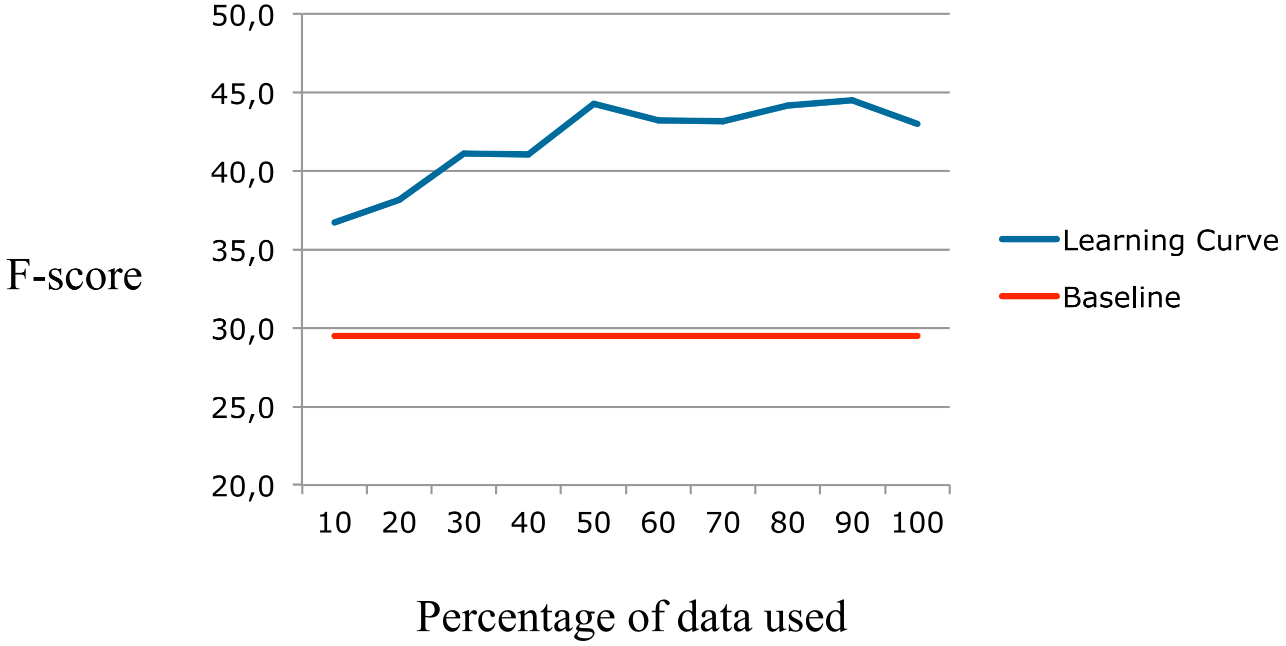
Influence of constituent

	Precision	Recall	F-score
Const 1	40.9	46.3	41.6
Const 2	39.3	42.7	38.7
Compound	45.2	48.4	45.6
Baseline	29.5		

- Larger influence of first constituent on the semantics of the compound (modifier)
- Similar to findings in psycholinguistics where first constituent has more influence on the selection of the linking element (Krott, Schreuder & Baayen, 2002)

Learning curves (1)

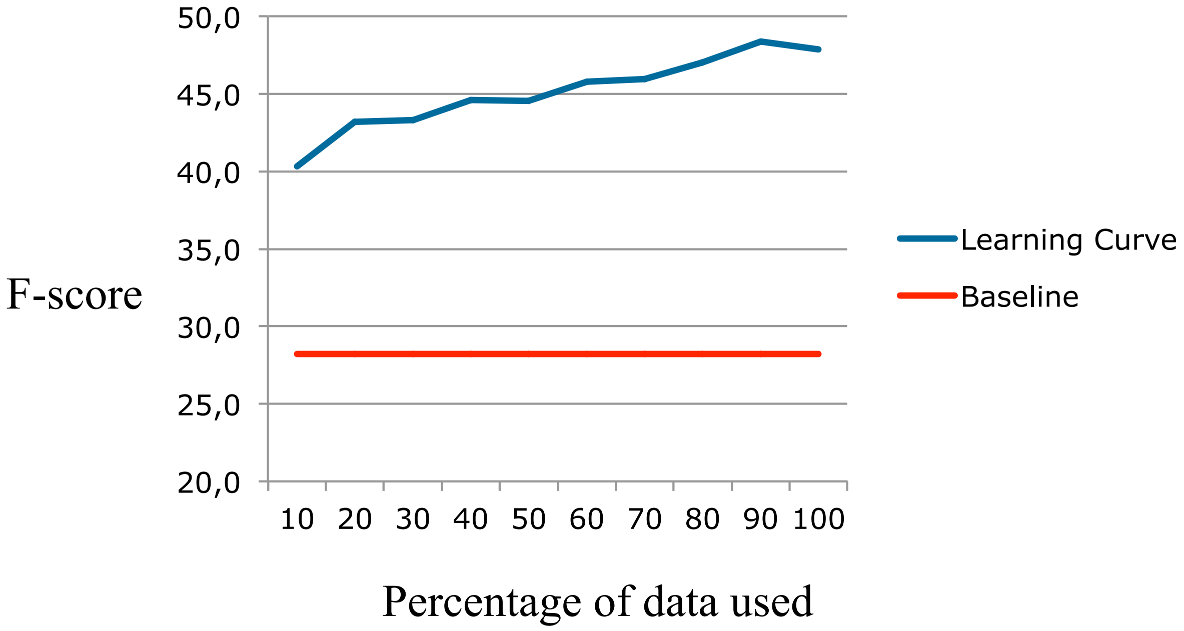
Dutch BOW 3



- Seems to quickly reach a ceiling
- Better than baseline

Learning curves (2)

Afrikaans BOW 3



- Seems somewhat more promising
- Yet, curve already starts high

- Either more systematicity in annotation
- Or slightly better corpus for this purpose

Discussion

- Is accuracy of 50% relevant?
  - Compare with human judgement: IAA of 50-60%.
  - Not all mistakes are stupid
    - Sometimes incorrect annotation and correct classification
      - E.g. *parochiestelsel* ‘parish system’
        - » Annotation: IN
        - » Classification: ABOUT
    - Sometimes both annotation and classification are correct
      - E.g. *badkuur* ‘bath treatment’
        - » Annotation: IN
        - » Classification: INST

Conclusion

- Promising initial results for both languages
- Highest F-scores
  - Afrikaans 51.1% (vs. 28.2%)
  - Dutch 47.3% (vs. 29.5%)
- Indication: Compares favourably with English research with similar methods
  - Ó Séaghdha 58.8%

Acknowledgement

Research sponsored by:

- Nederlandse Taalunie (Dutch Language Union)
- Departement of Arts and Culture (DAC) of South Africa
- National Research Foundation (NRF) of South Africa

Thank you!

For suggestions and/or questions:

Ben Verhoeven
CLiPS – Computational Linguistics Group
University of Antwerp
ben.verhoeven@uantwerpen.be

@clipsua