

# Mining Ontological Knowledge from Syntactically Annotated Corpora

Gosse Bouma  
Ismail Fahmi, Gertjan van Noord, Lonneke van der Plas

Information Science  
University of Groningen

LT4EL, Prague, June 07

# Outline

- 1 Introduction
  - Ontological Knowledge
  - Syntactic Analysis
- 2 Acquiring Ontological Knowledge
  - Term Extraction
  - Class Labels for Named Entities
  - Definition Sentences
  - Acronyms
  - Similar Words
  - Hypernyms
- 3 Conclusions

# Acquiring Ontological Knowledge

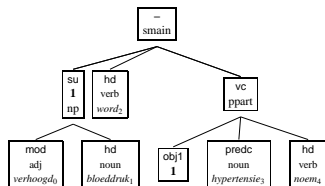
- **Acquiring Ontological Knowledge** from (Large) Corpora
  - Terms,
  - Definitions, acronyms, ...
  - Classes, ISA-relations, ..
  - Similar Words, Synonyms, ..
- How much **linguistic information** is needed?
  - Stems,
  - POS-labels, Named Entity Classes,
  - Chunks, grammatical relations, ...

# Syntactic Analysis

- Syntactic Analysis of Dutch: **Alpino**
- **Wide-coverage, robust, parser**
  - Hand-written feature-based grammar
  - Morphological analysis
  - Part-of-Speech tagger
  - Named Entity tagger
  - Maximum Entropy disambiguation model
- Produces **dependency relations**

## Dependency Relations

Verhoogde bloeddruk wordt hypertensie genoemd  
*High blood pressure is called hypertension*



<blood-pressure, mod, high>  
<is, su, blood-pressure>  
<is, vc, call>  
<call, predc, hypertension>  
<call, obj, blood-pressure>

## Processing large corpora

			M Words
<i>Automatically parsed</i>	TwNC	newspaper	500
	Wikipedia	web	50
	Europarl	proceedings	28
	IMIX med	various	3
	Name	Type	Accuracy
<i>Evaluation</i>	Trouw	newspaper	91.1
	CLEF	questions	96.3

## Term Extraction

Hypertension, commonly referred to as "high blood pressure", is a medical condition in which the blood pressure is chronically elevated. While it is formally called arterial hypertension, the word "hypertension" without a qualifier usually refers to arterial hypertension. Hypertension has been associated with a higher risk of heart attack or stroke.

- Which expressions are *terms*?

## Term Extraction

**Hypertension**, commonly referred to as "**high blood pressure**", is a medical condition in which the **blood pressure** is chronically elevated. While it is formally called **arterial hypertension**, the word "**hypertension**" without a qualifier usually refers to **arterial hypertension**. **Hypertension** has been associated with a higher risk of **heart attack** or **stroke**.

## Candidate multi-word terms

- **Corpus**: 1M words medical text (encyclopedia, hand book)
- Syntactic POS-filter
  - A\* N (Prep Det? A\* N) \*
  - 86K candidate multi-word terms extracted
- Rank candidate phrases by  $\chi^2$ , **Dice**, **MI**, log likelihood, t-test, frequency, ...
- **Evaluation**
  - **Gold Standard**: 29K terms obtained from ICD-9, wikipedia.nl, gezondheid.nl, encyclopedia index, ...
  - $\chi^2$ , **Dice**, **MI** give the most accurate results

## Improving Accuracy

- **Unithood**:  $\chi^2$  score
- **Termhood**: proportion of the words (stems) in the candidate term that can be found in a **list of known terms**
- **Evaluation**
  - Manual
  - Uninterpolated Average Precision (UAP)

$\alpha$	UAP
$\chi^2$ (baseline)	0.71
+ matching words	0.73
+ matching stems	0.84

## Changing the linguistic filter

- Use full parse instead of POS-filter
- Extract all NPs, strip only initial determiners and adverbs
- Include coordinations, proper PP -attachment, ..

	POS	Alpino
candidates	86K	96K
true terms	5.433	5.281

## Answering general WH-questions

- In which museum was the exhibition on Mondriaan?
  - `which(museum)`
- Which dynasty did Gengis Khan belong to?
  - `which(dynasty)`
- Name an evolution biologist.
  - `which(evolution biologist)`
- **NE** is a potential answer to **which(Concept)** if
  - **NE ISA Concept**

## (Short) Descriptions of Named Entities

- Who is Benazir Bhutto?
  - *Prime minister of Pakistan*
- What is Trans-Dniestr?
  - *unrecognized separatist republic inside Moldova*
- Who is Valentin Ivanov
  - *Russian referee*
- return most frequent **class label**
- expand with **modifiers** extracted from sentences where label was found

## Acquiring class labels for Named Entities

- Corpus is searched exhaustively for  $\langle$  **Concept**, app, Instance  $\rangle$ 
  - **museum** Hermitage, Madame Tussaud, National Gallery, ...
  - **Argentinian** newspaper La Nación, biologist Lilian Ramos, supermarket Disco, ....
  - peninsula: Al Faw (Iranian), Baja California (Mexican), Cape Cod, Jaffna, ...
- Noise (mostly attachment errors):
  - the biggest city of the **country**, **Rotterdam**
  - **language** of the country, **the Netherlands**

# Acquiring class labels for Named Entities

- **Corpus**: wikipedia.nl (50M words), newspapers (500M words)
- approx 5M pairs (2.1 M unique pairs) extracted
- Filter using **relative frequency**
  - $rf(\text{concept}, \text{NE}) = \text{freq}(\langle \text{Concept}, \text{app}, \text{NE} \rangle) / \text{freq}(\text{NE})$
  - filter all pairs with  $rf < 0.05$
- 392K unique pairs after applying filtering

## Effect of using Class labels

CLEF 2005 questions				
question type	baseline		improved	
	# q	score	# q	score
wh-questions	36	0.31	35	0.46
definition	60	0.53	60	0.68
person	26	0.69	7	0.71
function	0	0.00	20	0.75

# Concept Definitions

- What is leukopenia?
  - Leukopenia is a shortage of white blood cells (less than  $4 \times 10^9$  per liter)...
- What is kabuki?
  - Kabuki is a traditional Japanese theatre form which was developed during the Edo period.
- What is a cincinnato?
  - A cincinnato now is someone who has withdrawn himself to the country

# Definition Sentences

- Potential ⟨concept,definition⟩ sentences:
  - Contain a **copular use of the verb *zijn* (to be)**, and a nominal predicative complement
  - **Concept**: Head of the subject
  - **Definition**: predicative phrase
- **ok**
  - **Levitation** (from Latin *levare*, to raise) is **the process by which an object is suspended against gravity...**
  - **Döner Kebab** (...) is **the name given to a Turkish dish made with lamb (or mutton), beef or chicken.**
- **wrong**
  - A well-known **event** is **the fire in the village in 1641.**
  - His last **g**ame is **a fantasy RPG**

## Learning to Identify Definition Sentences

- Relevant **syntactic features**: presence of a determiner on subj/predc, definiteness of subj/predc, NEC of subj/predc, position of subj.
- Annotated 2500 potential definition sentences from **Wikipedia**
- Trained a (MaxEnt) classifier which distinguishes definitions from non-definitions, **using bigrams, syntactic features, sentence position**

Method	Precision
Baseline	59.4
Sentence position	75.9
<b>MaxEnt</b>	<b>92.2</b>

# Meaning of Acronyms

bij de **American Broadcasting Company** (ABC) ...  
De **Aanbevolen dagelijkse hoeveelheid** (ADH) ...  
... in een **medezeggenschapsraad** (MR)  
... de **Conferentie over Veiligheid en Samenwerking in Europa** (OVSE).

- Why use Syntax?
  - Acronym is **apposition** to a nominal phrase
  - **Preceding words in the phrase** must match (upper case initial, lower case internal), upper case need not match, lower case words may be skipped.

# Acronym Evaluation

- **Student systems:**
  - Developed on 40K sentences from Wikipedia with 2 upper case letters in sequence
  - Using pattern matching only
- Schwartz and Hearst (2003), *A simple algorithm for identifying abbreviation definitions in biomedical text*
- **Gold Standard** for Evaluation:
  - 1000 random sentences from development set
  - **Pooled** output of 9 student systems,
  - Removed errors

## Acronym Evaluation

fscore	precision	recall	rank	system
0.828	0.841	0.816	1	combined
0.774	0.892	0.684	2	student
0.770	0.868	0.691	3	schwartz/hearst
0.683	0.838	0.576	6	qatar2
0.621	0.645	0.598	8	qatar

**combined**: merged output of 4 stud sys with prec > 0.82

**qatar2**: qatar without non-matching upper case pattern

# Semantically Similar Words

- Members of the same semantic class (co-hyponyms)
  - apple, pear, orange (**fruit**)
  - BMW, Renault, Fiat (**cars**)
  - Java, C++, Perl, Ruby (**programming languages**)
- **Hypothesis:**
  - semantically similar words occur in syntactically similar contexts

## Co-occurrence data

- **Corpus**: wikipedia.nl (50M words), newspapers (500M words)

Relation	Example	Tuple	Size (M)
mod	betaalbare woning	Adj-N	21
object	woning renoveren	V-N	18
subject	woning vervuild	V-N	37
apposition	president Aristide	N-NE	12
Prep. Compl.	wisselen van woning	V+P-N	6
Coordination	woning en winkel	N-N	8

## Mutual Information

- **Mutual Information** for two words  $W_1, W_2$

$$MI = \log \frac{\text{freq}(W_1, W_2)}{\text{freq}(W_1) \times \text{freq}(W_2)}$$

	bouw	verlaat	koop	verkoop	plant	kap	heb	zie
huis	3.65	3.41	3.03	2.14	0.78	-	0.00	0.00
woning	4.50	3.03	2.19	2.09	-	-	0.00	0.00
boom	-	-	1.62	0.36	6.84	6.77	0.00	1.16
struik	-	-	-	-	7.02	5.94	0.00	0.72

## Comparing Vectors

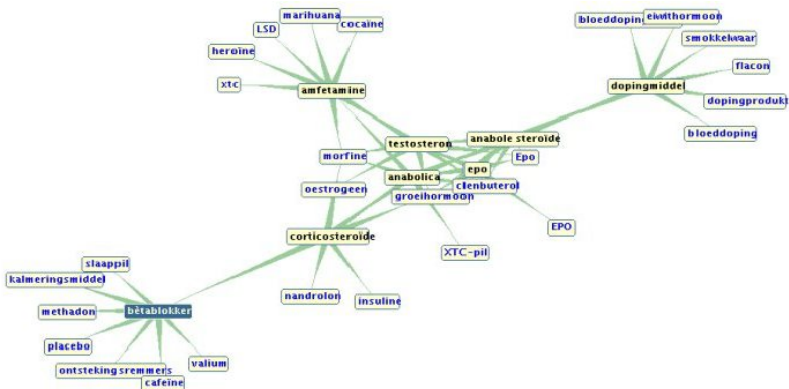
- Two words are similar if they have similar feature vectors
- DICE measure for comparing two vectors A and B

$$DICE = \frac{\text{Overlap}}{\text{Total}}$$

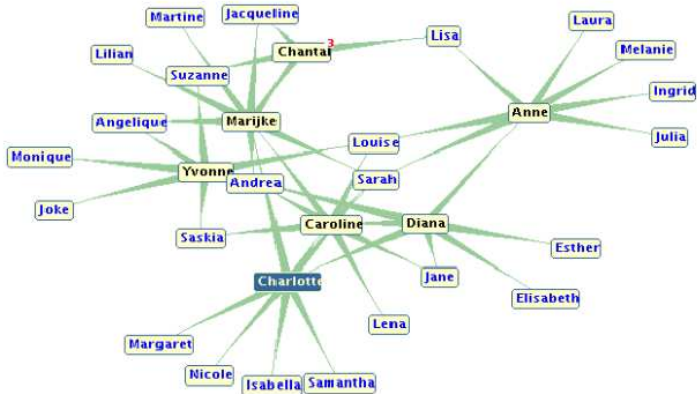
**Overlap** The sum of the average of all cells where A and B have a value  $> 0$

**Total** The sum of the average for all cells in A and B

# Similar Words



# Similar Words



## Lexical Acquisition

Noun	Apposition		Frequency
Diana	prinses	<i>princess</i>	154
Diana	vrouw	<i>wife</i>	6
Caroline	prinses	<i>princess</i>	1
Caroline	moeder	<i>mother</i>	1
Marijke	prinses	<i>princess</i>	4
Marijke	vriendin	<i>female friend</i>	3
Marijke	dochter	<i>daughter</i>	2
Marijke	moeder	<i>mother</i>	1
Yvonne	vrouw	<i>wife</i>	2
Yvonne	vriendin	<i>female friend</i>	1
Yvonne	moeder	<i>mother</i>	1
Diana	vrachtvaarder	<i>coaster</i>	2

# Learning Hypernym/Hyponym pairs

- **Hearst-patterns**
  - X and other Y
  - an X is a Y
  - Y, such as X
- high precision, but low recall

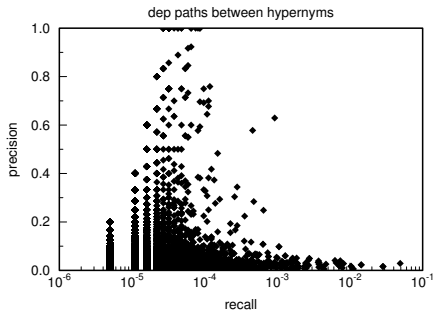
## Snow et al 2005

- Train a **classifier** using a large syntactically annotated corpus
- **Corpus Preparation**
  - Sentences containing two nouns  $N_1$  and  $N_2$  present in WordNet
  - Positive instance:  $N_1$  is a hypernym of  $N_2$
  - Negative instance:  $N_1$  is not a hypernym of  $N_2$
  - Feature: the syntactic dependency path from  $N_1$  to  $N_2$
- **Data collection:**
  - For each negative or positive pair  $N_1, N_2$ : count how often it occurs with path  $P$ .
- **Classifier** predicts whether a given pair is a hypernym/hyponym pair.

# Experiments

- Dutch **EuroWordNet** for hypernym/hyponym pairs
- 500M word parsed newspaper corpus
- Data construction:
  - Use 40K **most frequent paths** as features
  - Minimum pair frequency: 50
  - 575K pairs found (**566K** negative, **9K** positive)

# Results



MaxEnt, bin feats		
F-score	Prec	Rec
0.18	0.23	0.15

# Conclusions

- Some tasks require **hardly (no) linguistic information**
  - Find the meaning of acronyms
- Many tasks require **modest** access to linguistic information
  - Class labels for NE's, Definitions, ...
- Some tasks benefit from access to **syntactic dependencies**
  - Similar words, Hypernym/hyponym pairs,..