Closing Session

HS Experimentelles Arbeiten in der Sprachverarbeitung

Nils Reiter
nils.reiter@uni-koeln.de

January 26, 2023

Introduction

Two own experiments

- Nils Reiter/Anette Frank (2010). "Identifying Generic Noun Phrases". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jan Hajič/Sandra Carberry/Stephen Clark/Joakim Nivre. Uppsala, Sweden: Association for Computational Linguistics, pp. 40–49. URL: http://www.aclweb.org/anthology/P10-1005
  - Original slides from 2010!
- Benjamin Krautter/Janis Pagel/Nils Reiter/Marcus Willand (2020). "»[E]in Vater, dächte ich, ist doch immer ein Vater«. Figurentypen und ihre Operationalisierung". In: *Zeitschrift für digitale Geisteswissenschaften* 5. DOI: 10.17175/2020_007

Section 1

“Identifying Generic Noun Phrases”

Identifying Generic Expressions

Nils Reiter and Anette Frank

Department of Computational Linguistics
Heidelberg University
Germany

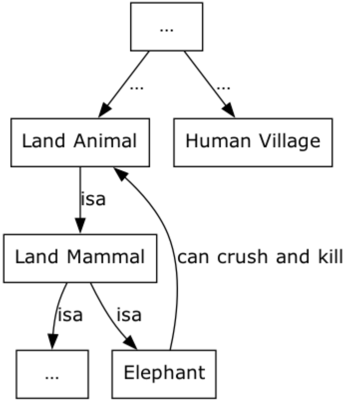
Elephants

[Elephants] can crush and kill any other land animal [...] In Africa, groups of young teenage elephants attacked human villages after cullings done in the 1970s and 80s.

?

Knowledge Acquisition

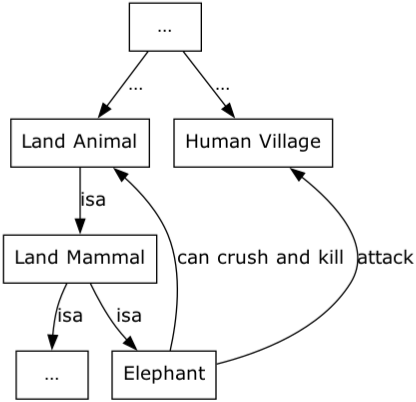
Elephants can crush and kill any other land animal. Groups of teenage elephants attacked human villages.



Hearst (1992), Cimiano (2006), Bos (2009)

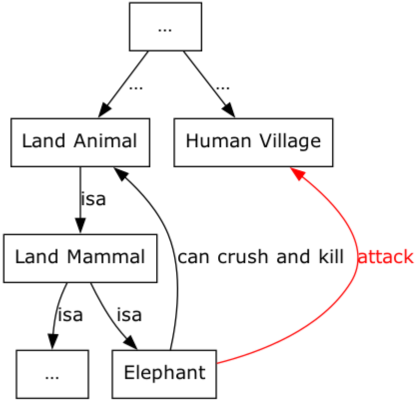
Knowledge Acquisition

Elephants can crush and kill any other land animal. Groups of teenage elephants attacked human villages.



Knowledge Acquisition

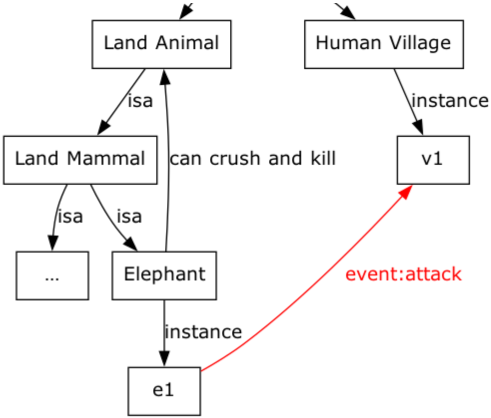
Elephants can crush and kill any other land animal. Groups of teenage elephants attacked human villages.



This is not a property of the class Elephant!

Knowledge Acquisition

Elephants can crush and kill any other land animal. Groups of teenage elephants attacked human villages.



It is a property of an instance of the class Elephant!

Starting Point

Knowledge acquisition systems need to be able
to distinguish classes and instances, otherwise

- Instance-level information is generalized to the class or
- Class-level knowledge is attached to instances

Starting Point

Knowledge acquisition systems need to be able
to distinguish classes and instances, otherwise

- Instance-level information is generalized to the class or
- Class-level knowledge is attached to instances

$\Rightarrow$ Identify generic noun phrases

Outline

Outline

Generic Noun Phrases

- Refer to a kind or class of individuals

Examples

- The lion was the most widespread animal.
- Lions eat up to 30 kg in one sitting.

Krifka et al. (1995)

Generic Sentences

- Express rule-like knowledge about habitual actions
- Do not express a particular event

Examples

- After 1971 [he] also took amphetamines.
- Lions eat up to 30 kg in one sitting.

Krifka et al. (1995)

Co-Occurrence

Example

Lions eat up to 30 kg in one sitting.

- This is a generic sentence that contains a generic noun phrase
- Both phenomena can (but don't have to) co-occur in a single sentence

Interpretations of Generic Noun Phrases

Quantification

- Quantification over individuals
- Exact determination of the quantifier restriction is extremely difficult
- Quantification over “relevant” or “normal” individuals

Dahl (1975), Declerck (1991), Cohen (1999)

Kind-Referring

- A generic NP refers to a kind
- Kinds are individuals that have properties on their own

Carlson (1977)

Interpretation of Generic Sentences

$$Q[x_1, ..., x_i](\underbrace{[x_1, ..., x_i]}_{\textit{Restrictor}}; \underbrace{\exists y_1, ..., y_i[x_1, .., x_i, y_1, ..., y_i]}_{\textit{Matrix}})$$

- Dyadic operator Q relates restrictor and matrix
- Generic operator quantifies over situations and events
- Exact determination of the quantifier restriction is extremely difficult

Heim (1982), Krifka et al. (1995)

Interpretation of Generic Sentences

$$Q[x_1, ..., x_i](\underbrace{[x_1, ..., x_i]}_{Restrictor}; \underbrace{\exists y_1, ..., y_i[x_1, .., x_i, y_1, ..., y_i]}_{Matrix})$$

- Dyadic operator Q relates restrictor and matrix
- Generic operator quantifies over situations and events
- Exact determination of the quantifier restriction is extremely difficult

Heim (1982), Krifka et al. (1995)

- Classification of generic sentences   Mathew and Katz (2009)

Characteristics

- No linguistic form of generic expressions

Examples (Noun Phrases)

- The lion was the most widespread mammal.
- A lioness is weaker [...] than a male.
- Elephants can crush and kill any other land animal.

Examples (Sentences)

- John walks to work.
- John walked to work (when he lived in California).
- John will walk to work (when he moves to California).

Outline

Aim

- Separate generic NPs from specific NPs
- Most of the tests and criteria given in the literature can't be operationalised
- Phenomena are context-sensitive

Aim

- Separate generic NPs from specific NPs
- Most of the tests and criteria given in the literature can't be operationalised
- Phenomena are context-sensitive

$\Rightarrow$ Corpus-based approach to identify generic noun phrases

Features

	Syntactic	Semantic
NP-level	Number, Person, Part of Speech, Determiner Type, Bare Plural	Countability, Granularity, Sense[0-3, Top]
S-level	Clause.{Part of Speech, Passive, Number of Modifiers}, Dependency Relation[0-4], Clause.Adjunct.{Verbal Type, Adverbial Type}, XLE.Quality	Clause.{Tense, Progressive, Perfective, Mood, Pred, Has temporal Modifier}, Clause.Adjunct.{Time, Pred}, Embedding Predicate.Pred

Table: Feature Classes

Feature Selection

Feature Combinations

- Each triple, pair and single feature tested in isolation

Ablation Testing

1. A single feature in turn is removed from the feature set
2. The feature whose omission causes the biggest drop in f-score is considered a strong feature
3. Remove strong feature and start over

In the end, we have a list of features sorted by their impact

Experiment: Corpus and Algorithm

Corpus

- ACE-2 corpus Mitchell et al. (2003)
- Newspaper texts
- 40,106 annotated entities
- 5,303 (13.2 %) marked as generic
- Balancing training data: $\sim$ 10,000 entities for each class
  - Over-sampling generic entities
  - Under-sampling non-generic entities

Experiment: Corpus and Algorithm

Corpus

- ACE-2 corpus Mitchell et al. (2003)
- Newspaper texts
- 40,106 annotated entities
- 5,303 (13.2 %) marked as generic
- Balancing training data: $\sim$ 10,000 entities for each class
  - Over-sampling generic entities
  - Under-sampling non-generic entities

Bayesian Network

- Weka implementation of a Bayesian net ?
- A Bayesian network represents dependencies between random variables as graph edges

Outline

Results of Feature Selection

Feature groups – singles, pairs, triples

- Most high ranking features are syntactic NP-level features (Number, POS, . . .)
- Few semantic features (Sense, Clause.{Tense, Pred})

Results of Feature Selection

Feature groups – singles, pairs, triples

- Most high ranking features are syntactic NP-level features (Number, POS, . . .)
- Few semantic features (Sense, Clause.{Tense, Pred})

Ablation Testing

- Clause-related features and dependency relations appear more often (and earlier) in the ablation results

Results of Feature Selection – Ablation

	Syntactic	Semantic
NP-level	Number, Person, Part of Speech, Determiner Type, Bare Plural	Countability, Granularity, Sense[0], Sense[1-3, Top]
S-level	Clause.Part of Speech, Clause.{Passive, Number of Modifiers}, Dependency Relation[2], Dependency Relation[0-1,3-4], Clause.Adjunct.{Verbal Type, Adverbial Type}, XLE.Quality	Clause.{Tense, Pred}, Clause.{Progressive, Perfective, Mood, Has temporal Modifier}, Clause.Adjunct.{Time, Pred}, Embedding Predicate.Pred

Table: Feature Classes

Baselines

Majority Each entity is non-generic

Person Use the feature Person

Suh Results of a pattern-based approach on detection of generic NPs Suh (2006)

	Generic			Overall		
	P	R	F	P	R	F
Majority	0	0	0	75.3	86.8	80.6
Person	60.5	10.2	17.5	84.3	87.2	85.7
Suh (2006)	28.9					

Table: Baseline results

Classification Results – Feature Classes

- Unbalanced data: syntactic features of the sentence and the NP perform best
- Balanced data: NP-syntactic features perform best
- All feature classes outperform baselines for the generic class, in terms of f-score

	Feature Set	Generic			Overall		
		P	R	F	P	R	F
	Baseline Person	60.5	10.2	17.5	84.3	87.2	85.7
Unbal.	Syntactic	40.1	66.6	50.1	87.2	82.4	84.7
	Semantic	34.5	56.0	42.7	84.9	80.1	82.4
	All	37.0	72.1	49.0	80.1	80.1	83.6
Balanced	NP/Syntactic	35.4	76.3	48.4	87.7	78.5	82.8
	S/Syntactic	23.1	77.1	35.6	85.1	63.1	72.5
	Syntactic	30.8	85.3	45.3	88.2	72.8	79.7
	Semantic	30.1	67.5	41.6	85.5	75.0	79.9
	All	33.7	81.0	47.6	88.0	76.5	81.8

Table: Classification results for some feature classes

Classification Results – Feature Selection

- Selecting features helps, results are better
- Ablation testing yields the feature set that outperforms every other feature set

	Feature Set	Generic			Overall		
		P	R	F	P	R	F
Baseline	Majority	0	0	0	75.3	86.8	80.6
	Person	60.5	10.2	17.5	84.3	87.2	85.7
	Suh (2006)	28.9					
Unbal.	5 best single features	49.5	37.4	42.6	85.3	86.7	86.0
	Feature groups	42.7	69.6	52.9	88.0	83.6	85.7
	Ablation set	45.7	64.8	53.6	87.9	85.2	86.5
Bal.	5 best single features	29.7	71.1	41.9	85.9	73.9	79.5
	Feature groups	35.9	83.1	50.1	88.7	78.2	83.1
	Ablation set	37.0	81.9	51.0	88.8	79.2	83.7

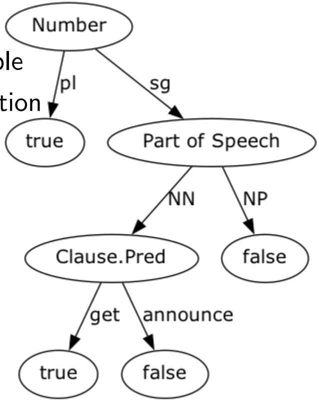
Table: Results of the classification for Feature Selection

Conclusions

- Corpus-based classification is feasible
- Features from all levels in combination perform best (Sentence vs. NP, Syntax vs. Semantics)
- Contextual factors with impact on the phenomenon can be uncovered

Conclusions

- Corpus-based classification is feasible
- Features from all levels in combination perform best (Sentence vs. NP, Syntax vs. Semantics)
- Contextual factors with impact on the phenomenon can be uncovered



Section 2

“»[E]in Vater, dächte ich, ist doch immer ein Vater«. Figurentypen und ihre Operationalisierung”

Computerlinguistik im B.A. Informationsverarbeitung

- Modul **Grundlagen der Computerlinguistik** (früher: Computerlinguistische Grundlagen)
  - Computerlinguistische Grundlagen (Seminar, Winter, Hermes)
    - Linguistische Grundlagen, Annotation
  - Sprachverarbeitung (Vorlesung + Übung, Sommer, Reiter)
    - Quantitative Eigenschaften von Sprache, Machine Learning
- Modul **Anwendungen der Computerlinguistik** (früher: Angewandte Linguistische Datenverarbeitung)
  - Deep Learning (Übung, Winter, Nester)
    - Deep Learning
  - Experimentelles Arbeiten in der Sprachverarbeitung (Hauptseminar, Winter, Reiter)
    - Experimente in der CL; wo kommen Fortschritt und Erkenntnis her?

Lernziele

- Lesen und verstehen NLP-technischer Forschungsliteratur
- Vertiefung vorhandener NLP-Kenntnisse
- Planung und Durchführung eigener Experimente

Closing

- Wie hat’s Ihnen gefallen?
- Was nehmen Sie mit?
- Was hat Ihnen gefehlt, was hätten Sie gerne mehr gemacht?


Gute Erholung!