Identifying Generic Expressions

Nils Reiter and Anette Frank

Department of Computational Linguistics
Heidelberg University
Germany

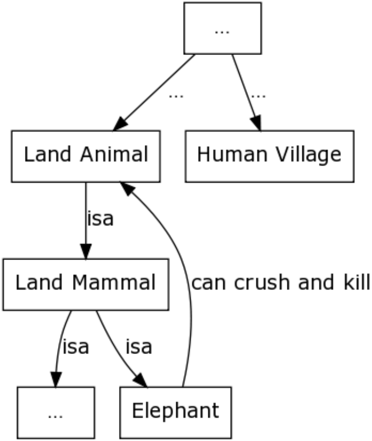
Elephants

[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.

Wikipedia (2010)

Knowledge Acquisition

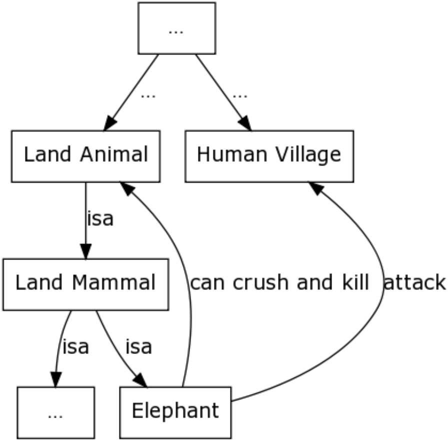
Elephants can crush and kill any other land animal.
Groups of teenage elephants attacked human villages.



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

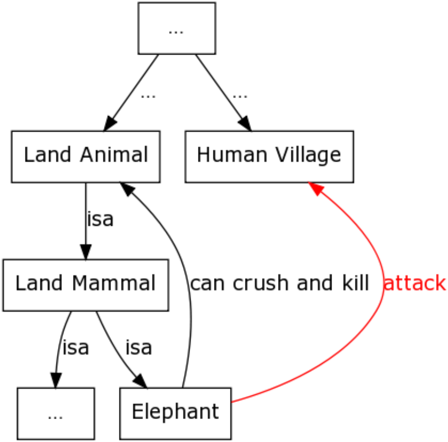
Knowledge Acquisition

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



Knowledge Acquisition

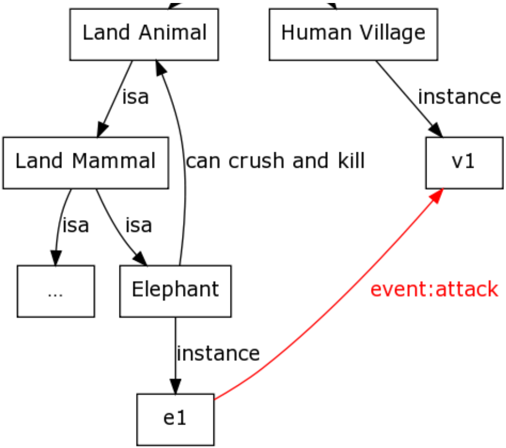
Elephants can crush and kill any other land animal.
Groups of teenage elephants attacked human villages.



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)

- <u>The lion</u> was the most widespread mammal.
- <u>A lioness</u> is weaker [...] than a male.
- <u>Elephants</u> 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 Witten and Frank (2002)
- 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

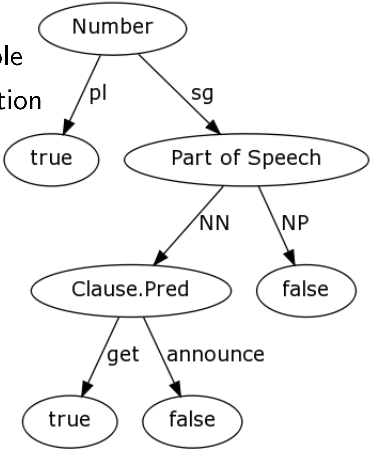
Table: Results of the classification for Feature Selection

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



Questions?

References I

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. CELEX2. Linguistic Data Consortium, Philadelphia, 1996.

Johan Bos. Applying automated deduction to natural language understanding. *Journal of Applied Logic*, 7(1):100 – 112, 2009.

Gregory Norman Carlson. *Reference to Kinds in English*. PhD thesis, University of Massachusetts, 1977.

Philipp Cimiano. *Ontology Learning and Populating from Text*. Springer, 2006.

Ariel Cohen. *Think Generic!: The Meaning and Use of Generic Sentences*. PhD thesis, Carnegie Mellon University, 1999.

Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. *XLE Documentation*, 2010. www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

References II

Östen Dahl. On Generics. In Edward Keenan, editor, *Formal Semantics of Natural Language*, pages 99–111. Cambridge University Press, Cambridge, 1975.

Renaat Declerck. The Origins of Genericity. *Linguistics*, 29: 79–102, 1991.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

Irene Heim. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts, Amherst, 1982.

Dan Klein and Christopher Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

References III

Manfred Krifka, Francis Jeffry Pelletier, Gregory N. Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. Genericity: An Introduction. In Gregory Norman Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*. University of Chicago Press, Chicago, 1995.

Thomas Mathew and Graham Katz. Supervised Categorization of Habitual and Episodic Sentences. In *Sixth Midwest Computational Linguistics Colloquium*. Bloomington, Indiana: Indiana University, 2009.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. ACE-2 Version 1.0. Linguistic Data Consortium, Philadelphia, 2003.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the conference on New Methods in Language Processing*, 12, 1994.

References IV

Sangweon Suh. Extracting Generic Statements for the Semantic Web. Master's thesis, University of Edinburgh, 2006.

Wikipedia. Elephant, 2010. URL `http://en.wikipedia.org/w/index.php?title=Elephant&direction=next&oldid=370885096`.

Ian H. Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record*, 31(1):76–77, 2002.

Results of Feature Selection

	Single	Pair	Triple
1	Bare Plural	Number, POS	Number, Clause.Tense, POS
2	Person	Countability, POS	Number, Clause.Tense, Noun type
3	Sense	Sense, POS	Number, Clause.POS, POS
4	Clause.Pred	Number, Countability	Number, POS, Noun type
5	EP.Pred	Noun type, POS	Number, Clause.POS, Noun type

Table: Best ranked features

Preprocessing

Task	Tool	
Sentence splitting	MorphAdorner [1]	
POS, lemmatization	TreeTagger	Schmid (1994)
WSD	MFS (according to WordNet 3.0)	
Countability	Celex	Baayen et al. (1996)
Parsing	XLE	Crouch et al. (2010)
	Stanford	Klein and Manning (2003)

Table: Preprocessing components

[1] http://morphadorner.northwestern.edu

Derived Feature Sets

Name	Description	Features
Set 1	Five best single features	Bare Plural, Person, Sense [0], Clause.Pred, Embedding Predicate.Pred
Set 2	Five best feature tuples	a. Number, Part of Speech b. Countability, Part of Speech c. Sense [0], Part of Speech d. Number, Countability e. Noun Type, Part of Speech
Set 3	Five best feature triples	a. Number, Clause.Tense, Part of Speech b. Number, Clause.Tense, Noun Type c. Number, Clause.Part of Speech, Part of Speech d. Number, Part of Speech, Noun Type e. Number, Clause.Part of Speech, Noun Type
Set 4	Features, that appear most often among the single, tuple and triple tests	Number, Noun Type, Part of Speech, Clause.Tense, Clause.Part of Speech, Clause.Pred, Embedding Predicate.Pred, Person, Sense [0], Sense [1], Sense[2]
Set 5	Features performing best in the ablation test	Number, Person, Clause.Part of Speech, Clause.Pred, Embedding Predicate.Pred, Clause.Tense, Determiner Type, Part of Speech, Bare Plural, Dependency Relation [2], Sense [0]

Table: Derived Features Sets

Classification Results – Feature Classes

Feature Set			Generic			Non generic			Overall		
			P	R	F	P	R	F	P	R	F
Baselines		Majority	0	0	0	86.8	100	92.9	75.3	86.8	80.6
		Person	60.5	10.2	17.5	87.9	99.0	93.1	84.3	87.2	85.7
		Suh (2006)	28.9								
Feature Classes	Unbalanced	NP	31.7	56.6	40.7	92.5	81.4	86.6	84.5	78.2	81.2
		S	32.2	50.7	39.4	91.8	83.7	87.6	83.9	79.4	81.6
		NP/Syntactic	39.2	58.4	46.9	93.2	86.2	89.5	86.0	82.5	84.2
		S/Syntactic	31.9	22.1	26.1	88.7	92.8	90.7	81.2	83.5	82.3
		NP/Semantic	28.2	53.5	36.9	91.8	79.2	85.0	83.4	75.8	79.4
		S/Semantic	32.1	36.6	34.2	90.1	88.2	89.2	82.5	81.4	81.9
		Syntactic	40.1	66.6	50.1	94.3	84.8	89.3	87.2	82.4	84.7
		Semantic	34.5	56.0	42.7	92.6	83.8	88.0	84.9	80.1	82.4
		All	37.0	72.1	49.0	81.3	87.6	87.4	80.1	80.1	83.6
	Balanced	NP	30.1	71.0	42.2	94.4	74.8	83.5	85.9	74.3	79.7
		S	26.9	73.1	39.3	94.4	69.8	80.3	85.5	70.2	77.1
		NP/Syntactic	35.4	76.3	48.4	95.6	78.8	86.4	87.7	78.5	82.8
		S/Syntactic	23.1	77.1	35.6	94.6	61.0	74.2	85.1	63.1	72.5
		NP/Semantic	24.7	60.0	35.0	92.2	72.1	80.9	83.3	70.5	76.4
		S/Semantic	26.4	66.3	37.7	93.3	71.8	81.2	84.5	71.1	77.2
		Syntactic	30.8	85.3	45.3	96.9	70.8	81.9	88.2	72.8	79.7
		Semantic	30.1	67.5	41.6	93.9	76.1	84.1	85.5	75.0	79.9
		All	33.7	81.0	47.6	96.3	75.8	84.8	88.0	76.5	81.8

Table: Results of the classification for Feature Classes

Classification Results – Feature Selection

Feature Set			Generic			Non generic			Overall		
			P	R	F	P	R	F	P	R	F
Baselines		Majority	0	0	0	86.8	100	92.9	75.3	86.8	80.6
		Person	60.5	10.2	17.5	87.9	99.0	93.1	84.3	87.2	85.7
		Suh (2006)	28.9								
Feature Selection	Unbalanced	Set 1	49.5	37.4	42.6	90.8	94.2	92.5	85.3	86.7	86.0
		Set 2a	37.3	42.7	39.8	91.1	89.1	90.1	84.0	82.9	83.5
		Set 3a	42.6	54.1	47.7	92.7	88.9	90.8	86.1	84.3	85.2
		Set 4	42.7	69.6	52.9	94.9	85.8	90.1	88.0	83.6	85.7
		Set 5	45.7	64.8	53.6	94.3	88.3	91.2	87.9	85.2	86.5
	Balanced	Set 1	29.7	71.1	41.9	94.4	74.4	83.2	85.9	73.9	79.5
		Set 2a	36.5	70.5	48.1	94.8	81.3	87.5	87.1	79.8	83.3
		Set 3a	36.2	70.8	47.9	94.8	81.0	87.4	87.1	79.7	83.2
		Set 4	35.9	83.1	50.1	96.8	77.4	86.0	88.7	78.2	83.1
		Set 5	37.0	81.9	51.0	96.6	78.7	86.8	88.8	79.2	83.7

Table: Results of the classification for Feature Selection