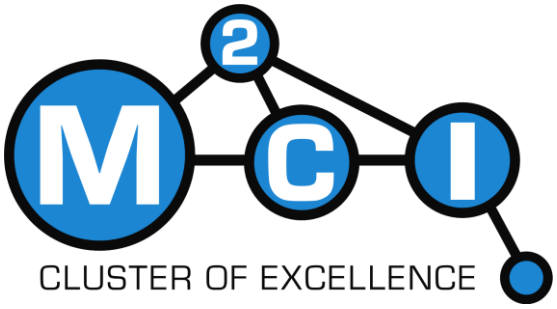
UNIVERSITÄT DES SAARLANDES



Discourse-sensitive Automatic Identification of Generic Expressions

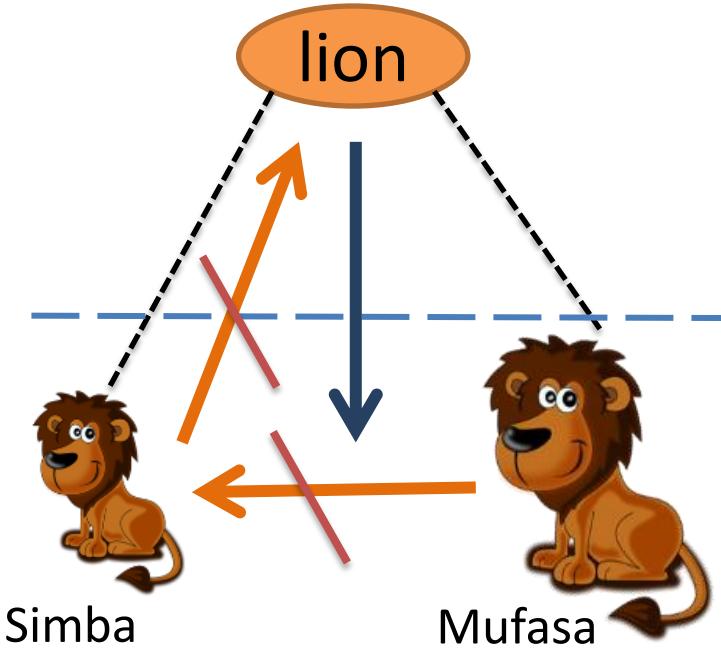
Annemarie Friedrich **Manfred Pinkal**

afried,pinkal@coli.uni-saarland.de

Computational Linguistics, Universität des Saarlandes

ACL 2015, BEIJING, CHINA.

Generic vs. non-generic expressions



different
entailment properties

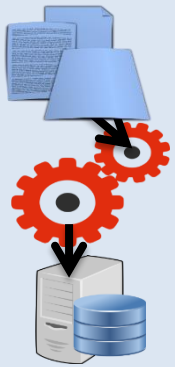
Lions are dangerous.

kind-referring generic

Mufasa is dangerous.
Simba is dangerous.

non-generic

Automatic identification:
why?



knowledge extraction from text

contribution of clauses to **discourse structure**:

characterizing statements
≠ particular events or states

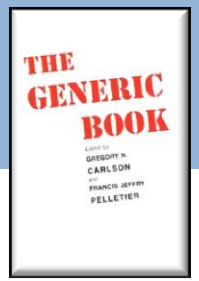
natural language understanding

How? Discourse context matters

WIKIPEDIA
The Free Encyclopedia

Sugar maples also have a tendency to color unevenly in fall. **generic**
The recent year's growth twigs are green and turn dark brown. **generic**

Discourse-sensitive approach:
sequence labeling task
classification of (subject) **noun phrases** & **clauses**



Reference to kinds

Krifka et al. (1995): Genericity: An Introduction.

	kind-referring	**non-kind-referring**
definite NPs	The lion is a predatory cat.	The cat chased the mouse.
indefinite NPs	Lions eat meat.	Dogs were barking outside.
quantified NPs	Some (type of) dinosaur is extinct.	Some dogs were barking outside.
proper names	Panthera leo persica was first described by the Austrian zoologist Meyer.	John likes ice cream.

form of NP not sufficient

clause / context matters

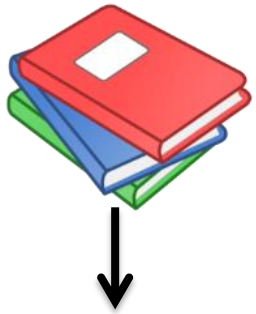
Annotation scheme

motivated by **semantic theory** (Krifka et al. 1995)
more details: Friedrich et al. (LAW 2015)

clause / subject	**generic** (characterizing statements about kinds)	**non-generic** (statements about particular things/people, particular events/states)
generic	*Lions have manes.* *Lions eat meat.*	*Dinosaurs died out.* *The blobfish was voted the "World's Ugliest Animal".*
non-generic	*does not exist by definition*	*John is a nice guy.*

Footnote for linguists: identification of habitual sentences is left to future work.

WikiGenerics corpus



102 Wikipedia texts

about animals, sports, politics, science, biographies, …

balanced corpus ~50% generic

10279 clauses

SPADE system
Soricut & Marcu (ACL 2003)

segmentation into clauses

Labels

subject \ clause	generic	non-generic
generic	GEN_gen	NONGEN_gen
non-generic		NONGEN_non-gen

majority vote

gold standard

Fleiss' κ

subject	clause	subject + clause
0.69	0.72	0.68

substantial agreement

Computational model

Sugar maples also have a tendency to color unevenly in fall.

The recent year's growth twigs are green.

sequence of clauses (entire document)

barePlural=true 1
determinerType=def : 0
tense=present : 1
voice=active : 1
…

barePlural=true : 0
determinerType=def : 1
…
currentLabel=GEN and previousLabel=GEN : 1
…

features: indicator functions

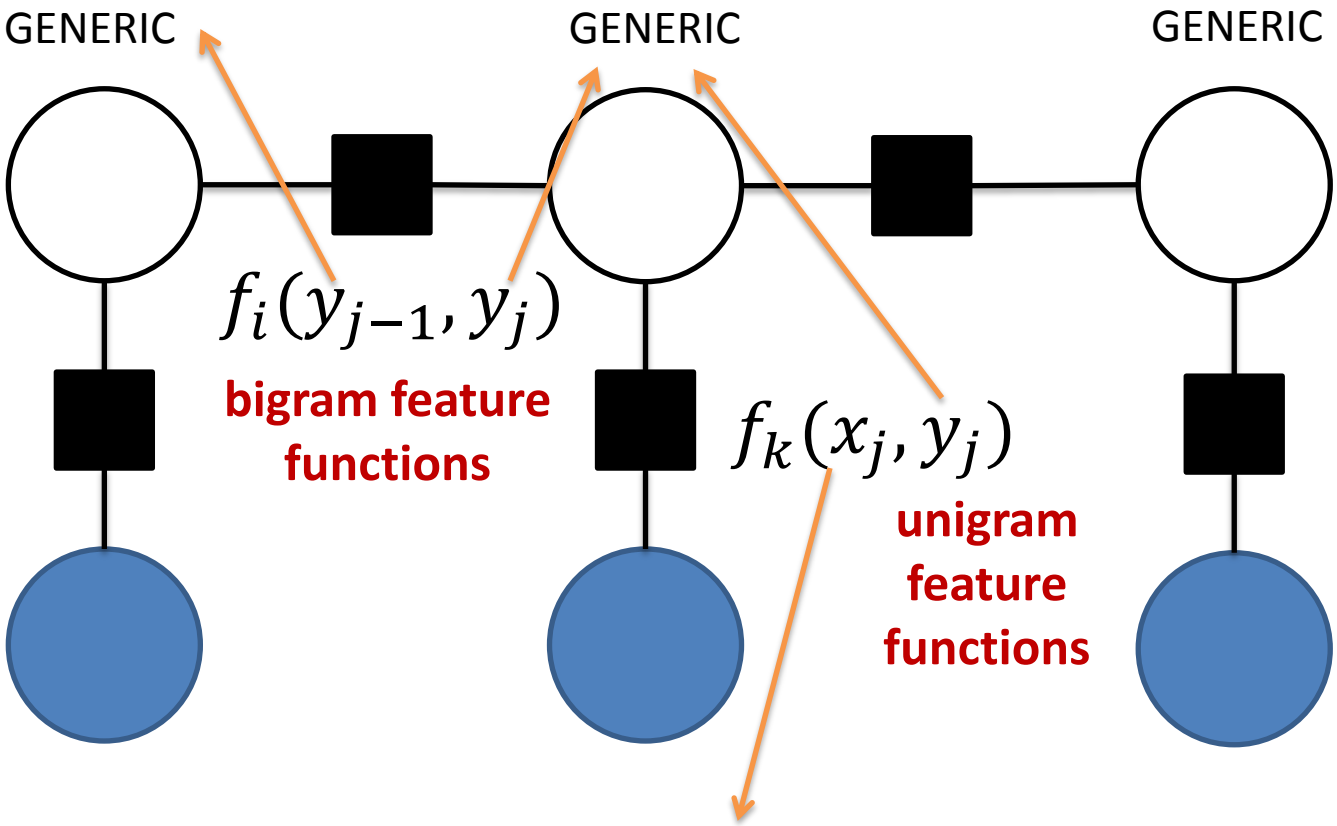
CRF

GENERIC

GENERIC

sequence of labels

Linear-chain Conditional Random Field



label sequence $\vec{y}$

$$P(\vec{y}|\vec{x})$$

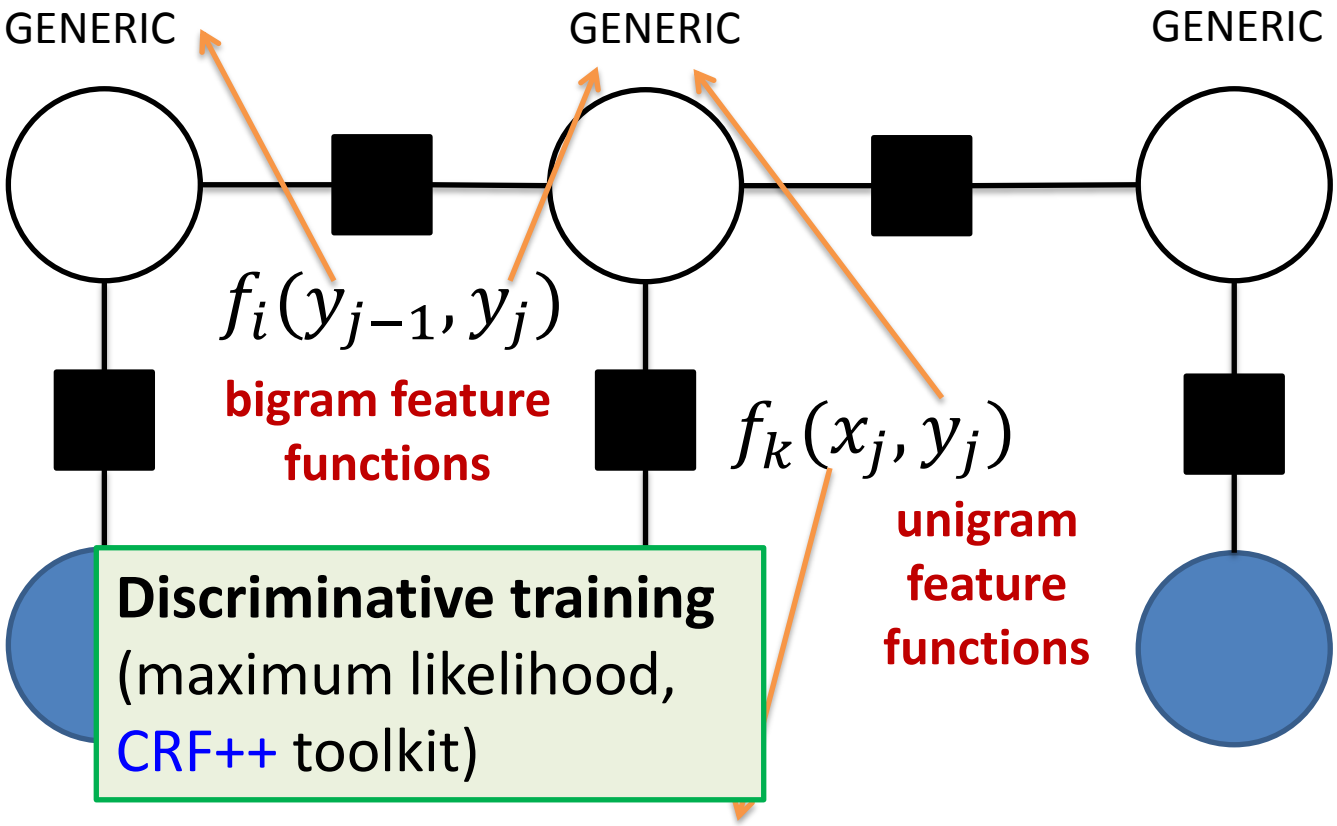
observation sequence $\vec{x}$

Acer saccharum is a deciduous tree.

Sugar maples also have a tendency to color unevenly in fall.

The recent year's growth twigs are green.

Linear-chain Conditional Random Field



label sequence $\vec{y}$

$$P(\vec{y}|\vec{x})$$

observation sequence $\vec{x}$

Acer saccharum is a deciduous tree.

Sugar maples also have a tendency to color unevenly in fall.

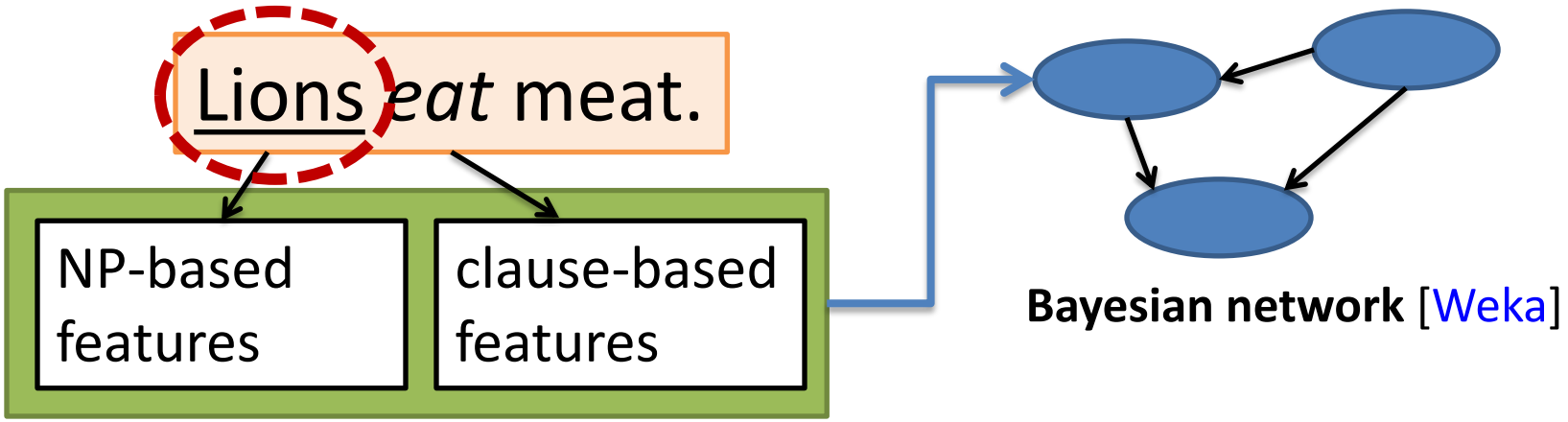
The recent year's growth twigs are green.

Baseline [Reiter & Frank (ACL 2010)]

Data: ACE-2 & ACE-2005

→ largest corpora annotated with NP-level genericity to date, ~40k NPs

→ SPC = specific / non-generic, GEN = generic, USP = underspecified



→ we use the same feature set for our CRF model

subject:	generic/non-generic
clause:	generic/non-generic
subject+clause:	GEN_gen, NONGEN_gen, NONGEN_non_gen

R&F baseline for clause / subject+clause tasks: BayesNet trained on our labels

Features [see Reiter & Frank (ACL 2010)]

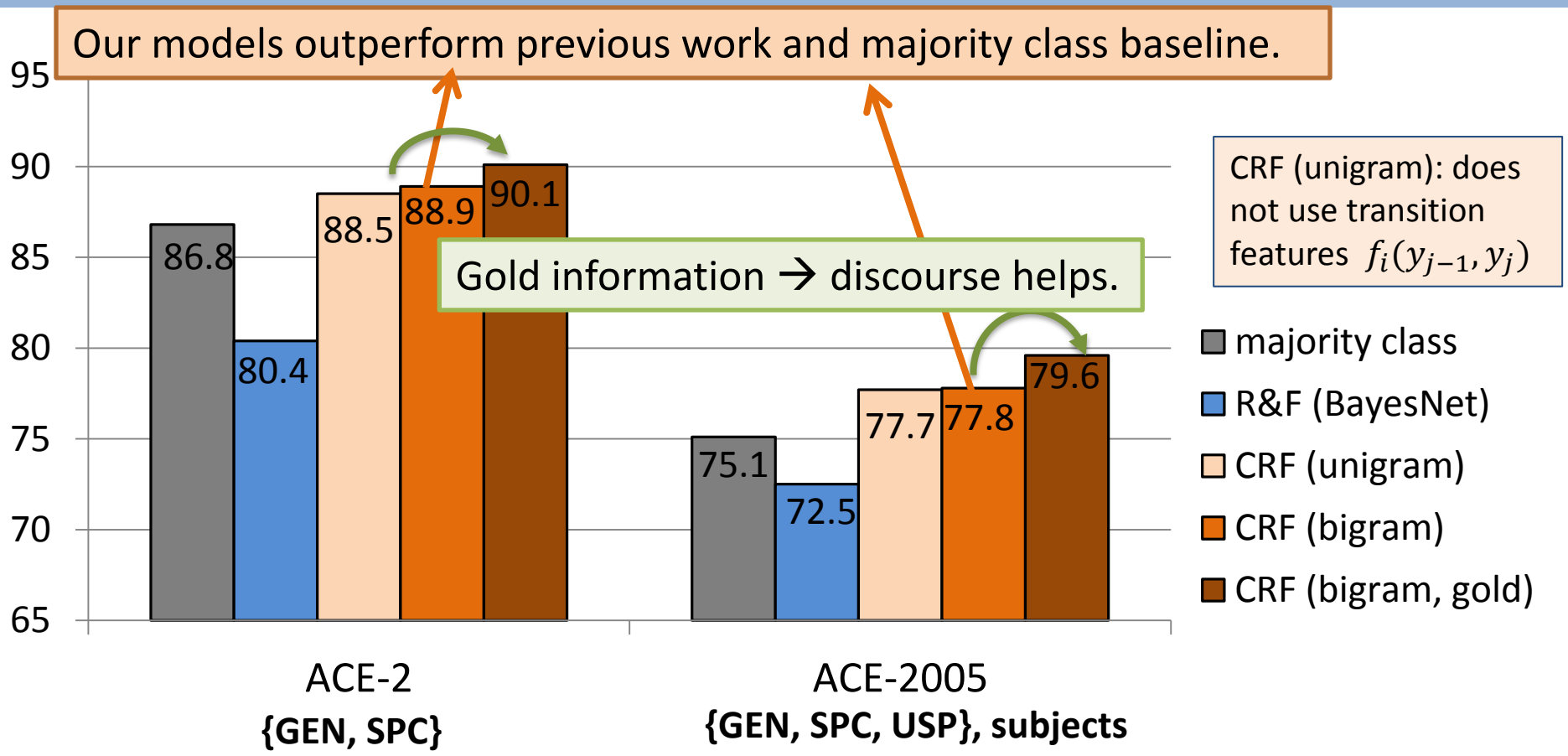
→ reimplementation of R&F using freely available resources.

→ extracted from dependency parses (Stanford parser)

NP-based features	
number	sg, pl
person	1,2,3
countability	Celex: count, uncount,…
noun type	common, proper, pronoun
determiner type	def, indef, demon
part-of-speech	POS of head
bare plural	true, false
WordNet based features	senses, lexical filename,…

Clause-based features	
dependency relations	between (subject) head and governor etc.
tense	past, present, future
progressive	true, false
perfective	true, false
voice	active, passive
part-of-speech	POS of head
temporal modifier	true, false
number of modifiers	numeric
predicate	lemma of head
adjunct-degree	positive, comparative, superlative

Accuracy: ACE-2 and ACE-2005

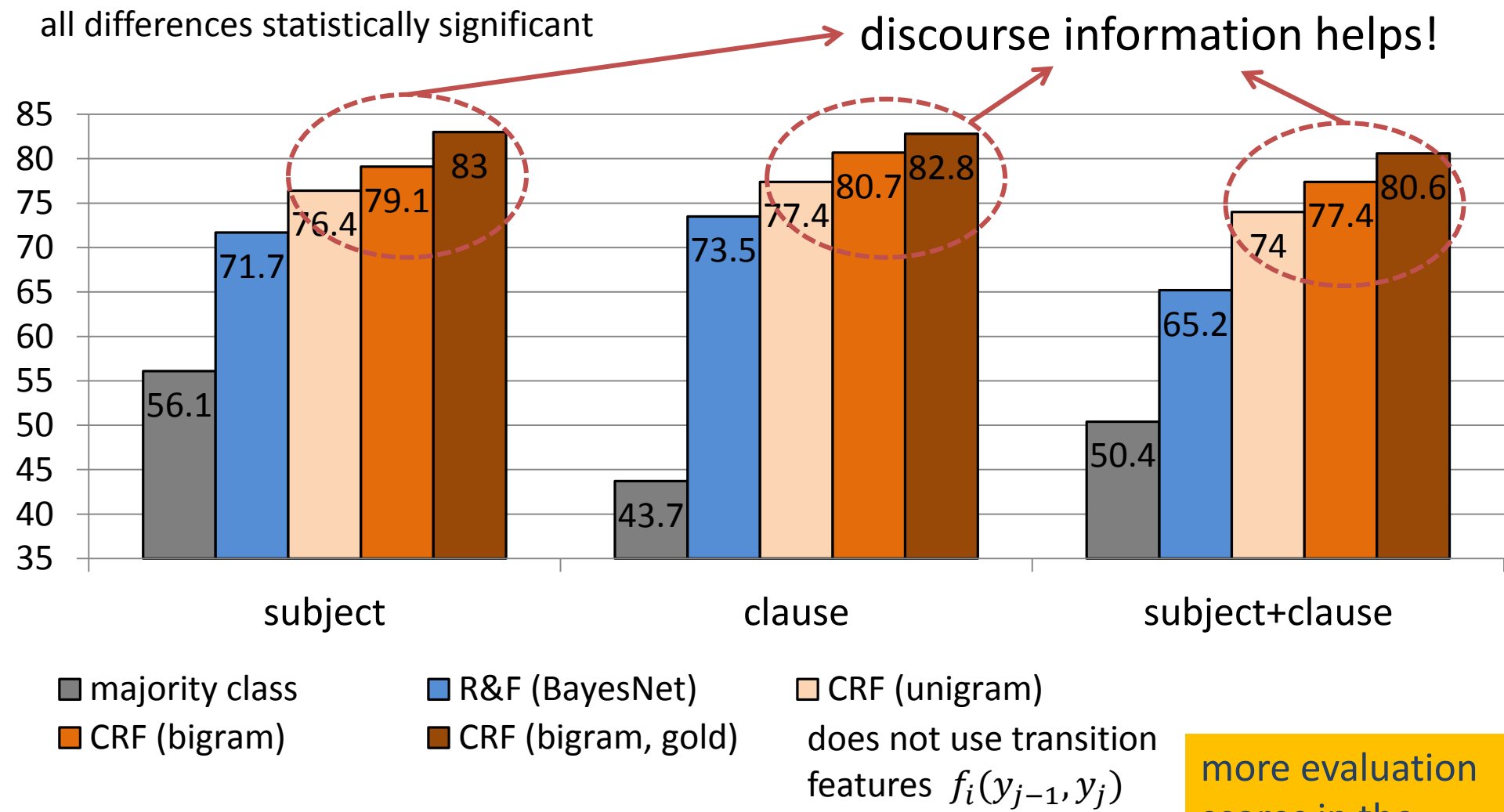


Few generic instances. (for details see Friedrich et al. (LAW 2015))

Problems in annotation guidelines, mix genericity and specificity.

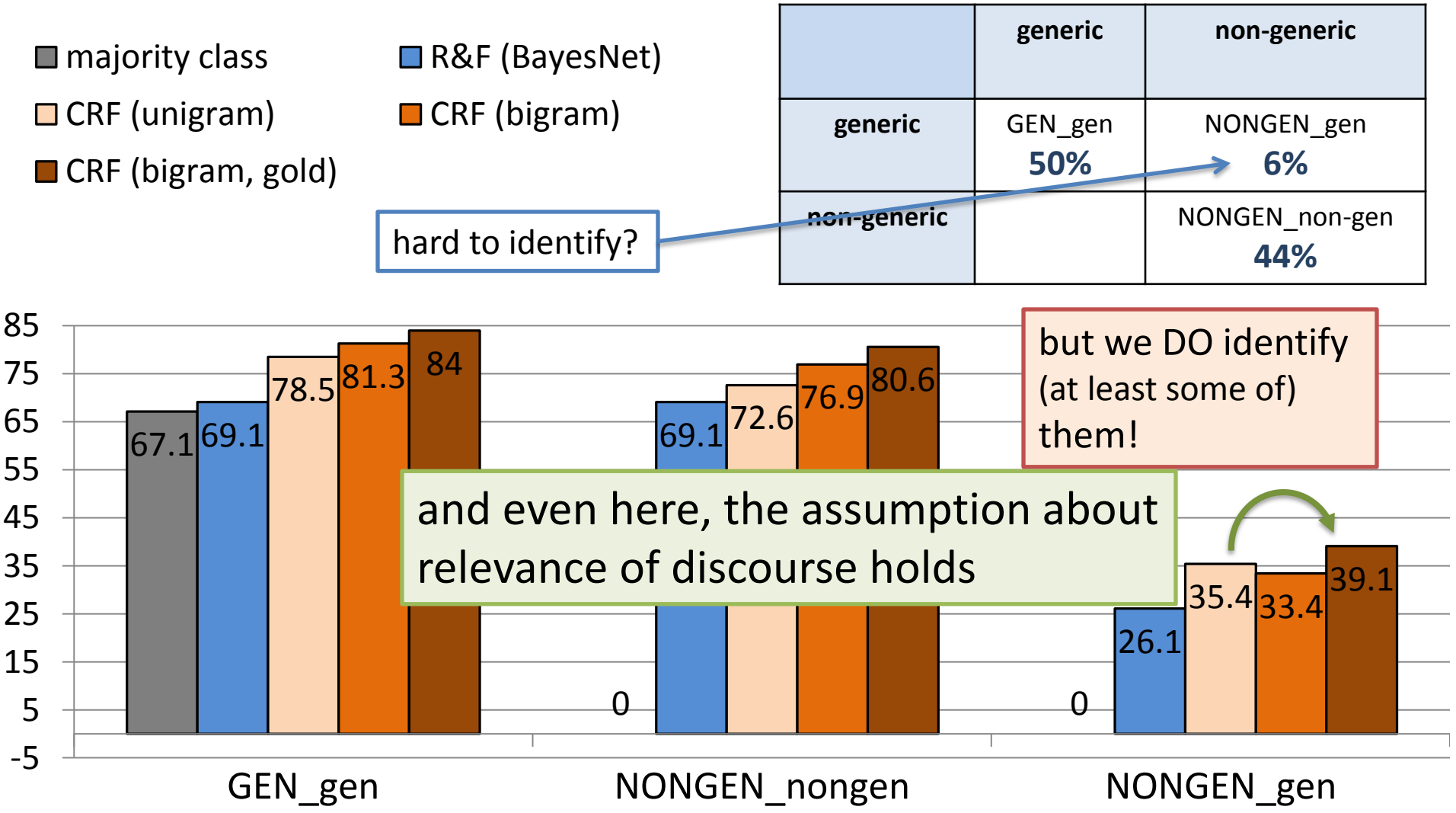
→ *Officials reported…* (USP) → is non-generic (SPC), non-specific!

Accuracy: WikiGenerics



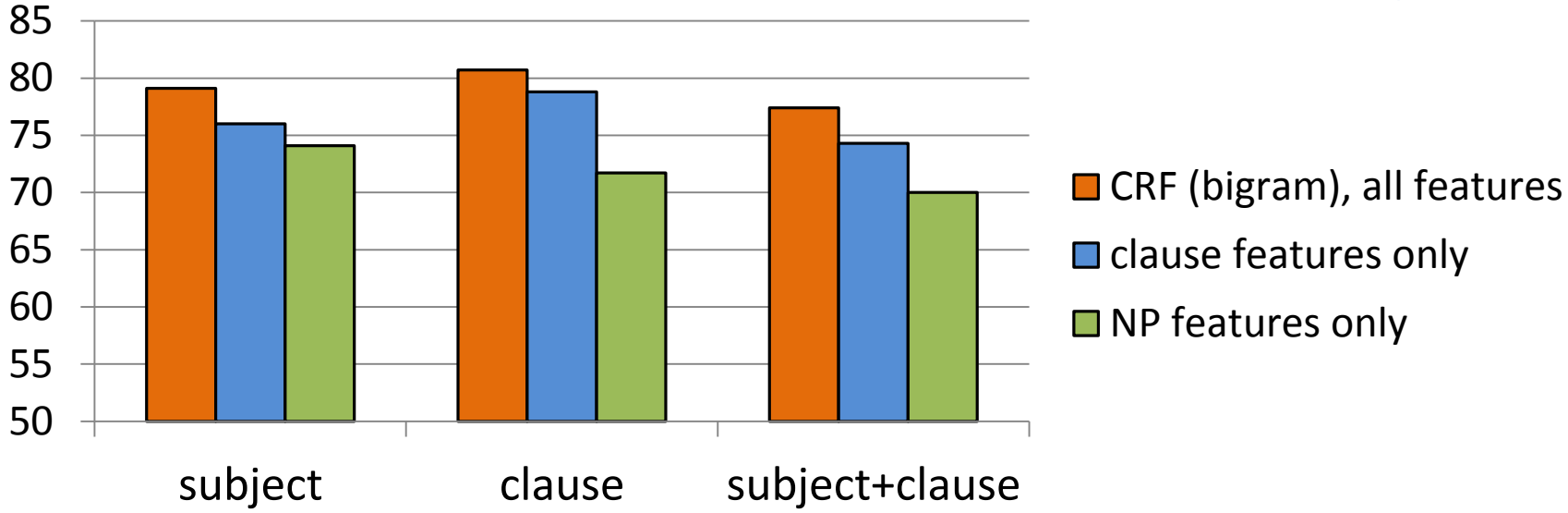
more evaluation scores in the paper!

F1-scores: subject + clause



Model inspection

Feature set ablation: NP or clause features more important?



It strongly depends on the **clause** whether an NP or a clause are interpreted as generic or not!

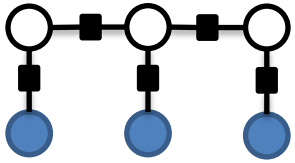
Markov order: integrate more preceding labels?

- no need to use higher orders, using only the preceding label is optimal
- labels of non-adjacent clauses ***do*** influence each (score is optimized for entire sequence)

Conclusions & future work

We classify **NPs** and **clauses** with regard to their genericity.

WikiGenerics corpus
balanced
substantial agreement



CRF finds **optimal label sequence** for clauses of a document, combining information from clause and surrounding labels

discourse information matters!

FUTURE WORK

Genericity of NPs other than the subject

Cats chase mice.

Related linguistic phenomena

John cycled to work today. (episodic)
John cycles to work. (habitual)

Data set & implementation of features:
www.coli.uni-saarland.de/projects/sitent

Thank you!

Special thanks to: Alexis Palmer, Melissa Peate Sørensen, Nils Reiter, Christine Bocionek and Kleo-Isidora Mavridou.

References

ACE corpora: https://www.ldc.upenn.edu/collaborations/past-projects/ace

Friedrich, A., Palmer, A., Peate Sorensen, M. & Pinkal, M. (2015). **Annotating genericity: a survey, a scheme, and a corpus.** In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, USA

Krifka, M. et al. (1995). **Genericity: an introduction**. *The Generic Book*, 1-124. University of Chicago Press.

Reiter, N., & Frank, A. (2010, July**). Identifying generic noun phrases.** In *Proceedings of ACL (*pp. 40-49). Association for Computational Linguistics.

Soricut, R., & Marcu, D. (2003). **Sentence level discourse parsing using syntactic and lexical information.** ACL-HLT. (pp. 149-156). Association for Computational Linguistics.

Comparison of CRF-bigram and CRF-unigram: Examples

This appendix contains some examples for cases that the CRF-unigram model got wrong, but the CRF-bigram model got right. In general, more gains are observed for the **non-generic** class. In the larger part of these cases, the bigram model seems to "make up" for missing coreference resolution, as in the following example, (cases that the unigram model gets wrong but the bigram model gets right are marked in blue):

The invention of the modern piano is credited to Bartolomeo Cristofori who was employed by Ferdinando de' Medici, Grand Prince of Tuscany, as the Keeper of the Instruments; he was an expert harpsichord maker. ***(non-generic)***

During the summer, narwhals mostly eat Arctic cod and Greenland halibut, with other fish such as polar cod making up the remainder of their diet. Each year, they migrate from bays into the ocean as summer comes. ***(generic)***

It is comparably easier to manually identify **generic** cases in the data that are correctly classified as **generic** by the bigram model, but which even humans could not classify correctly without seeing the discourse context. Here are some of the interesting examples.

A species popular among aquaculturists is the Piaractus mesopotamicus, also known as "Paraná River Pacu". Pacus inhabit most rivers and streams in the Amazon and Orinoco river basins of lowland Amazonia. "Some pacus? Or the kind pacu?" The blue sentence itself is underspecified, but the context indicates that the sentence talks about the kind Pacu. ***(generic)***

Archimedes' screw consists of a screw (a helical surface surrounding a central cylindrical shaft) inside a hollow pipe. The screw is turned usually by a windmill or by manual labour. As the shaft turns, the bottom end scoops up a volume of water. This water will slide up in the spiral tube, until it finally pours out from the top of the tube and feeds the irrigation systems. The screw was used mostly for draining water out of mines or other areas of low lying water. "The particular screw I'm holding in my hand?" The context indicates that the sentence talks about a type of screw. ***(generic)***

Grimpoteuthis is a genus of pelagic umbrella octopus that live in the deep sea. Prominent ear-like fins protrude from the mantle just above their lateral eyes. "Does this describe some particular individuals or does it refer to a kind?" ***(generic)***

The helpful context may also occur after the clause in question.
The study indicated that sloths sleep just under 10 hours a day. Three-toed sloths are mostly diurnal, while two-toed sloths are nocturnal. "The study" is non-generic here, but all other NPs are ***generic****.*

Shlemovidnye gusli is a variety of Gusli held by the musician on his knees, so that the strings are horizontal, the resonator body under them. He uses his left hand to mute unnecessary strings. Out of context, the blue sentence would rather sound like a non-generic one. However, here, 'he' refers to the hypothetical musician and is hence ***generic****, too. This is also a case of "making up for missing coreference resolution".*

In his sixth semester, Koch began to conduct research at the Physiological Institute, where he studied succinic acid secretion. This would eventually form the basis of his dissertation. *'This' refers to the particular research Koch did. Using the context, the bigram model makes a plausible decision to label this as non-generic here.*