FONDAZIONE
BRUNO KESSLER

Corpus Pattern Analysis for Textual Entailment

Vlad Niculae

FbK, Trento - Italy

August 10, 2012

Foreword

FONDAZIONE BRUNO KESSLER

Who Vlad is & why not to delete his e-mail

- Machine learning, scientific computing in Python guy

- Twice GSoC student for **scikit-learn**, sprint with them
- 173 followers on **Twitter** (`@vnfrombucharest`)
- 43 views per day on average on geeky blog `blog.vene.ro`
- Charismatic, gives good talks
- Looks good in group pictures
- Has 9 good fingers


SIN

Corpus Pattern Analysis for Textual Entailment

The RTE challenge

- **Text ($\mathcal{T}$):** For weeks now, I've been hearing chatter that Disney was close to doing a distribution deal with Hulu. Disney would give Hulu exclusive access to at least some of its online video in exchange for an equity stake alongside GE's (GE) NBC and News Corp.'s (NWS) Fox. And accompanying said chatter was this refrain: Why? The puzzlement comes from video players who don't work at NBC, Fox or Hulu, and who can't see the upside in Disney CEO Bob Iger throwing in his lot with Hulu.
- **Hypothesis ($\mathcal{H}$):** Bob Iger is the CEO of Disney.

Does $\mathcal{T}$ entail $\mathcal{H}$?

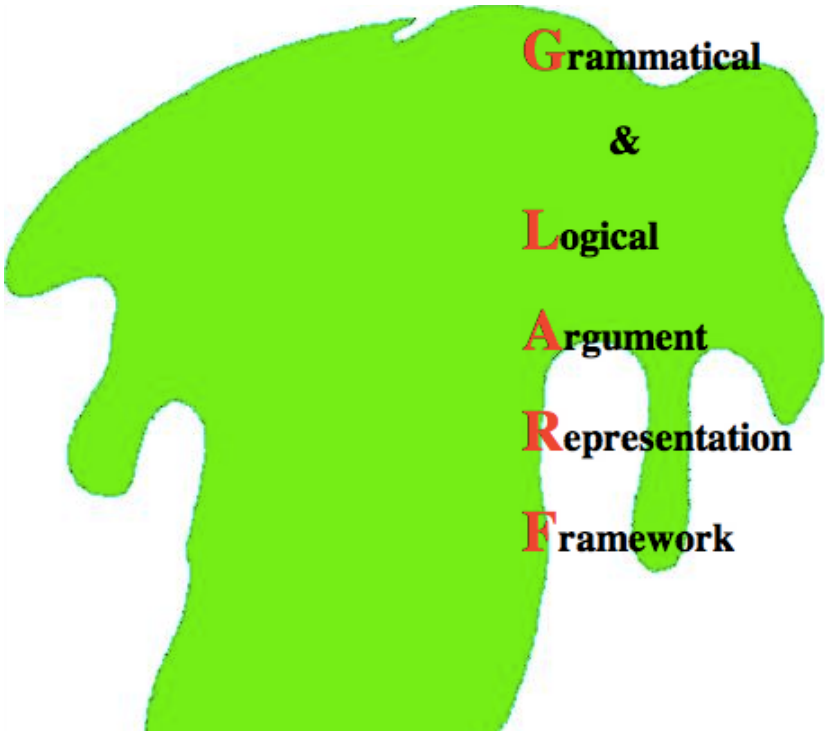
Would a human, after reading $\mathcal{T}$, assert that $\mathcal{H}$ is most probably true?

Successful approaches:

- Feed features (word overlap, tree overlap, NEs, negative words) to an SVM
- Most measure accuracy
- Precision and Recall are usually balanced. Should they be?
- What if? **Hypothesis:** Hulu is the CEO of Disney.

GLARF

`http://nlp.cs.nyu.edu/meyers/GLARF.html`



"GLARF is a typed feature structure framework for representing regularizations of parse trees." It is built on top of the Charniak parser and JET.

GLARF

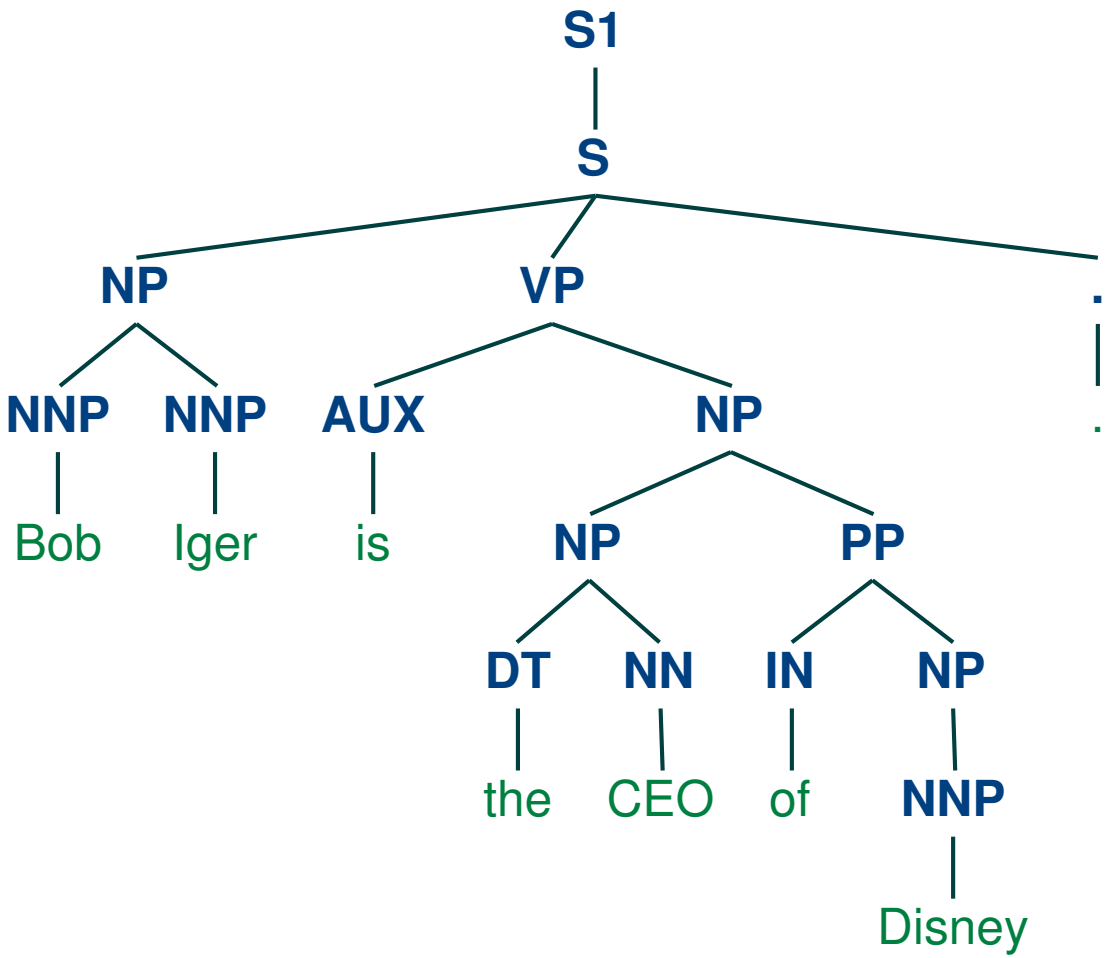
- Theoretical syntax (Zellig & Haris, LFG, HPSG) + CL resources (COMLEX, NOMLEX, PTB, PropBank)
- Relations are not only head-dependent
- 3 kinds of relations: Logic1 (DAG), Surface (tree) and Logic2 (DG)
- A complex treatment of clauses

Glarf does a lot of the "hidden work" we need

- NER integrated in the syntax: SEM, NE, NE-subtype multiword
- Normalization: time, figures, affiliated, title, name, post-hon
- Attribute disentanglement: apposite, relative, transparent

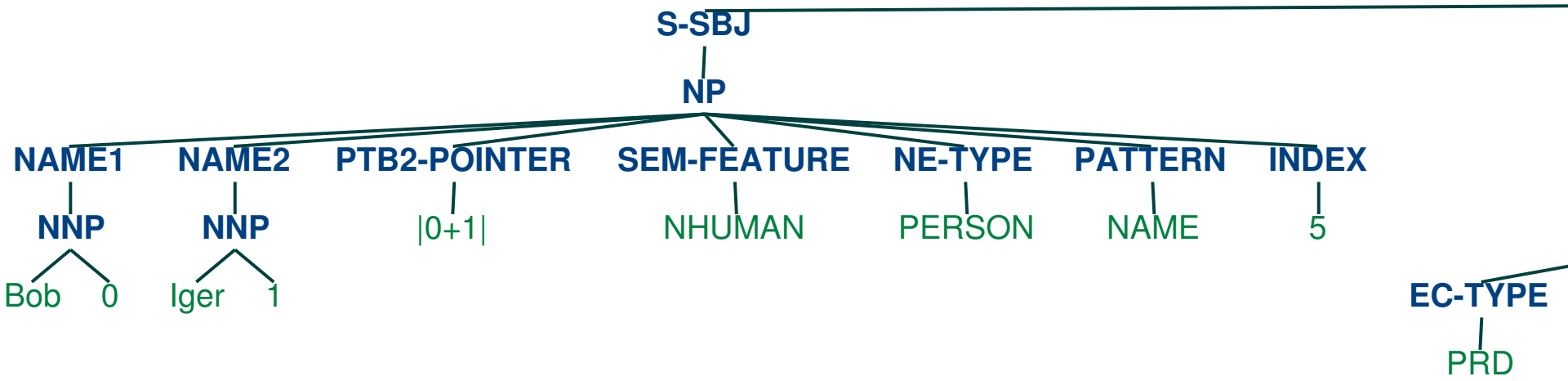
FONDAZIONE
BRUNO KESSLER

Parse tree

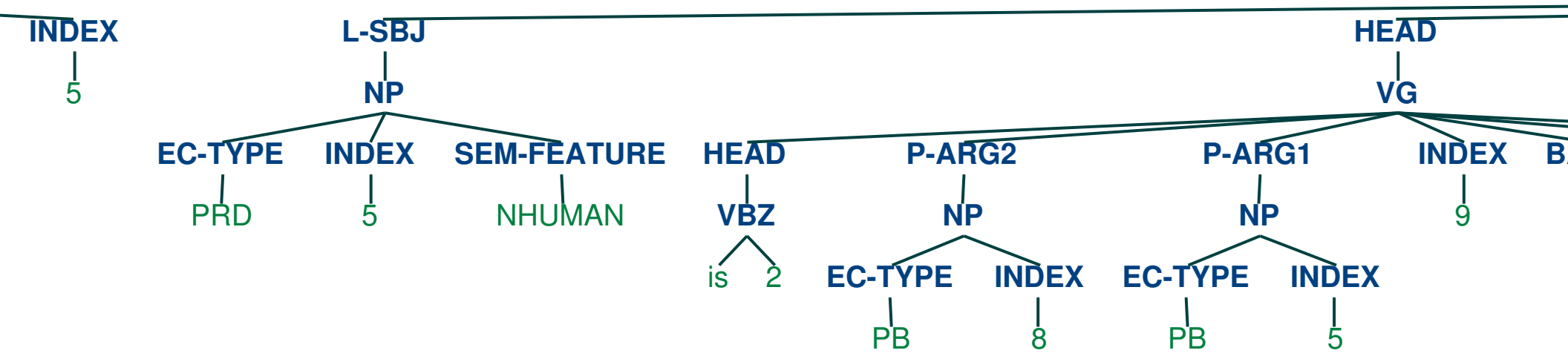


FONDAZIONE
BRUNO KESSLER

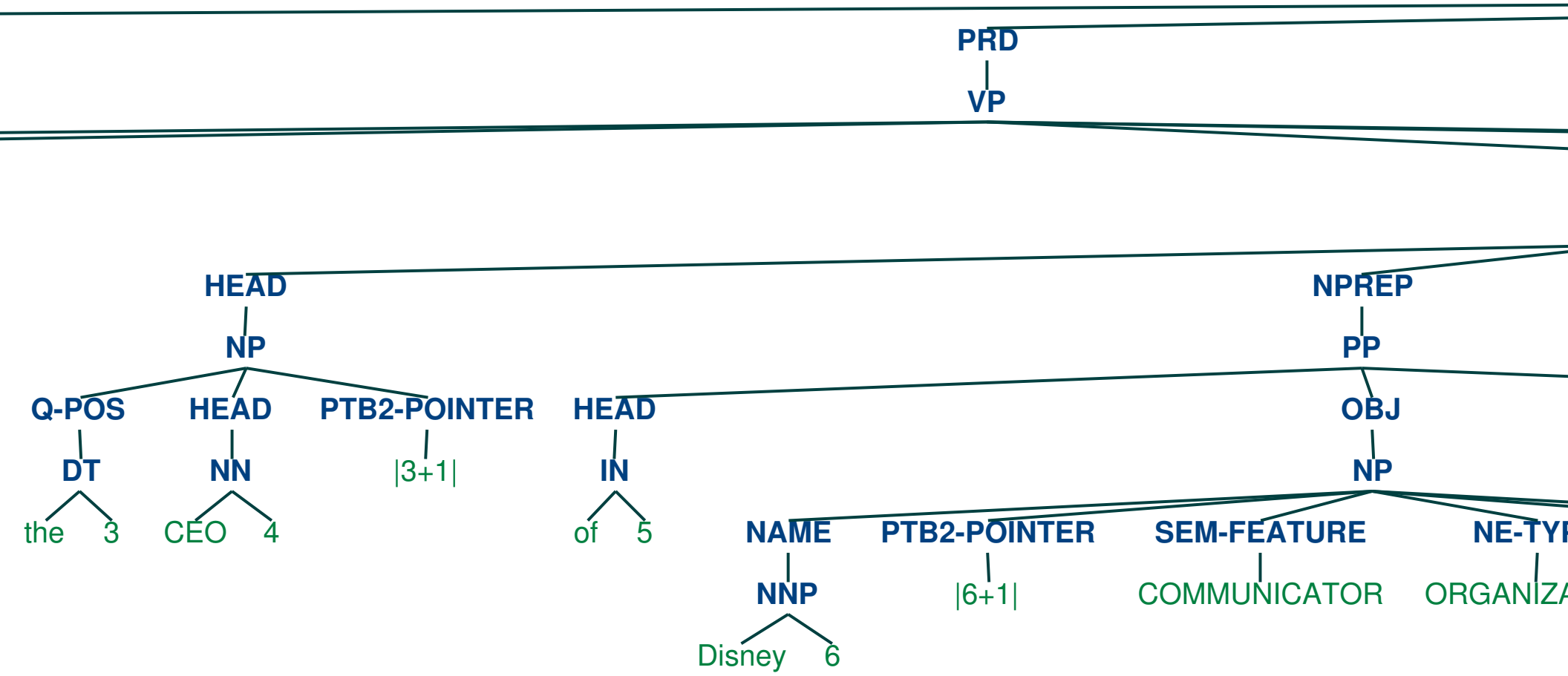
Glarfed tree



Glarfed tree (continued)



Glarfed tree (enough already!!)



How to deal with Glarfed trees

- **NLTK** `Tree` module can parse them (with appropriate tweaks)
- You could write code on top of that to interpret Glarf's own tags...
- Copy-paste driven workflow
- Why not call Glarf straight from Python?...

The answer

Use (and extend) **pyglarf** !
(Project page: `https://github.com/vene/pyglarf`)

FONDAZIONE BRUNO KESSLER

pyglarf makes your life easier

```
>>> with GlarfWrapper() as gw:
...     _, _, glarf_out = gw.make_sentences(
            "John died in Boston in 1972.")
...
>>> tree = GlarfTree.glarf_parse(glarf_out[0])
>>> for rel in tree.rels():
...     print rel
...
DIE/1 [CATEGORY: VG, INDEX: 10, SENSE-NAME:
       "STERBEN", VERB-SENSE: 1, PARENT_CATEGORY: VP]
P-ARG1 [SBJ NP INDEX: 5]:  John/0 (NP+NAME 0-0)
P-ARGM-LOC [ADV2 PP INDEX: 12]:  in/4 |1972|/5 (PP+HEAD 4-5)
ADV1 [PP INDEX: 11]: in/2 Boston/3 (HEAD+IN 2-2, OBJ+NP 3-3)
ADV2 [PP INDEX: 12]: in/4 |1972|/5 (HEAD+IN 4-4, OBJ+NP 5-5)
```

FONDAZIONE BRUNO KESSLER

A first step: ISA hypotheses

- Focus on only the pairs where the hypothesis has the form X is α
- (Not where "is" plays the role of AUX though.)
- Glarf regularizes these nicely across tenses, voices, etc.
- `Baseline 1:` Lexical overlap $\rightarrow$ SVM (close to Yashar, Magnini)
- `Baseline 2:` Use more features: (see `nltk.classify.rte_classify`)
  - # common words (that are not NEs)
  - # words exclusive to $\mathcal{H}$
  - # common NEs
  - # NEs exclusive to $\mathcal{H}$
  - # negative words in $\mathcal{T}$
  - # negative words in $\mathcal{H}$
  - # common constituents of Charniak parse trees

Recognizing ISA entailment

The gist of it:

1. Run $\mathcal{T}$ and $\mathcal{H}$ through Charniak and GLARF
2. Match the ISA pattern: $\mathcal{H} \sim X$ is α
3. Extract *entities* and their *attributes* from $\mathcal{T}$:
   $\mathcal{T} \sim X_1(\alpha_1, \alpha_2, \dots), \dots, X_n(\alpha_k, \dots)$
4. If an entity $X_i \equiv X$ can be found in $\mathcal{T}$ and $X_i(\alpha)$, then declare entailment.

Example

- **Text ($\mathcal{T}$):** For weeks now, I've been hearing chatter that Disney was close to doing a distribution deal with Hulu. Disney would give Hulu exclusive access to at least some of its online video in exchange for an equity stake alongside GE's (GE) NBC and News Corp.'s (NWS) Fox. And accompanying said chatter was this refrain: Why? The puzzlement comes from video players who don't work at NBC, Fox or Hulu, and who can't see the upside in Disney CEO Bob Iger throwing in his lot with Hulu.
- **Hypothesis ($\mathcal{H}$):** Bob Iger is the CEO of Disney.

Example

- **Text ($\mathcal{T}$):** For weeks now, I've been hearing chatter that Disney was close to doing a distribution deal with Hulu. Disney would give Hulu exclusive access to at least some of its online video in exchange for an equity stake alongside GE's (GE) NBC and News Corp.'s (NWS) Fox. And accompanying said chatter was this refrain: Why? The puzzlement comes from video players who don't work at NBC, Fox or Hulu, and who can't see the upside in Disney CEO Bob Iger throwing in his lot with Hulu.
- **Hypothesis ($\mathcal{H}$):** Bob Iger is the CEO of Disney.

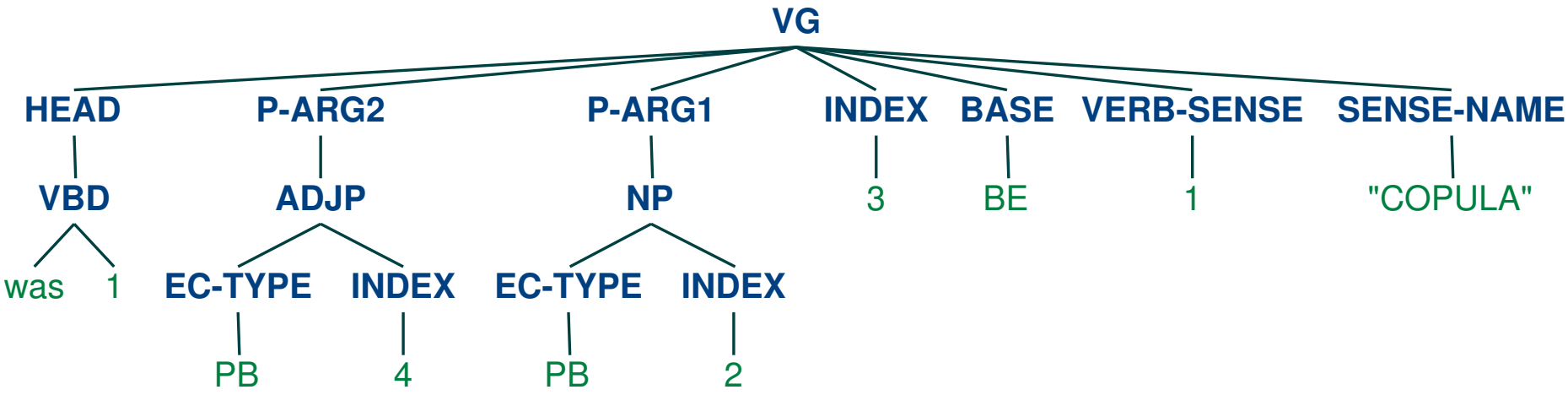
Example

- **Text ($\mathcal{T}$):** For weeks now, I've been hearing chatter that Disney was close to doing a distribution deal with Hulu. Disney would give Hulu exclusive access to at least some of its online video in exchange for an equity stake alongside GE's (GE) NBC and News Corp.'s (NWS) Fox. And accompanying said chatter was this refrain: Why? The puzzlement comes from video players who don't work at NBC, Fox or Hulu, and who can't see the upside in Disney CEO Bob Iger throwing in his lot with Hulu.
- **Hypothesis ($\mathcal{H}$):** Bob Iger is the CEO of Disney.

Example

- **Text ($\mathcal{T}$):** For weeks now, I've been hearing chatter that Disney was close to doing a distribution deal with Hulu. Disney would give Hulu exclusive access to at least some of its online video in exchange for an equity stake alongside GE's (GE) NBC and News Corp.'s (NWS) Fox. And accompanying said chatter was this refrain: Why? The puzzlement comes from video players who don't work at NBC, Fox or Hulu, and who can't see the upside in Disney CEO Bob Iger throwing in his lot with Hulu.
- **Hypothesis ($\mathcal{H}$):** Bob Iger is the CEO of Disney.

Matching the ISA pattern



- Account for *transparency*: An apple is a kind of fruit.
- Account for prepositions: Martin was in Canada.

FONDAZIONE BRUNO KESSLER

Extracting entities and attributes

- Disjoint set (union-find) data structure
- Every NP starts out as its own entity
- Join entities with the same HEAD or linked by APPOSITION / AFFILIATION
- Add pre-modifiers and complements as attributes
  (`NP` $\rightarrow$ `(X-POS)* HEAD (COMP)*`)
- Join entities and attach attributes from matched ISAs (**M**)

Representing entities:

- The title of an entity is the NP that Glarf tagged with the most semantic info
- Everything else becomes an attribute

Matching: *non-trivial!*

- Assume that string inclusion is entailment.
- Counterexample (that we get wrong): **Vice President**
- Reach out with synonyms (American Thesaurus, Roget's Thesaurus) (**S**)

FONDAZIONE BRUNO KESSLER

Results

# +/-	RTE3 104 / 91				RTE4 90 / 102				RTE5 134 / 131			
	A	P	R	F	A	P	R	F	A	P	R	F
BL 1	.69	.71	.69	.70	.51	.49	.35	.40	.57	.56	.84	.67
BL 2	.73	.69	.87	.77	.55	.52	.78	.62	.59	.59	.80	.68
*	.51	.83	.11	.19	.54	.73	.09	.16	.38	.79	.08	.15
S	.51	.70	.13	.23	.59	.71	.22	.34	.41	.68	.19	.30
SP	.51	.72	.12	.21	.60	.78	.20	.32	.40	.74	.15	.25
SPM	.54	.74	.22	.34	.61	.76	.24	.37	.40	.62	.19	.29
SP'	.55	.64	.38	.47	.61	.59	.52	.56	.47	.59	.43	.50
SPM'	.59	.70	.40	.51	.59	.56	.53	.55	.48	.59	.48	.53

Artificially generated samples

- **Text ($\mathcal{T}$):** For weeks now, I've been hearing chatter that Disney was close to doing a distribution deal with Hulu. Disney would give Hulu exclusive access to at least some of its online video in exchange for an equity stake alongside GE's (GE) NBC and News Corp.'s (NWS) Fox. And accompanying said chatter was this refrain: Why? The puzzlement comes from video players who don't work at NBC, Fox or Hulu, and who can't see the upside in Disney CEO Bob Iger throwing in his lot with Hulu.
- **Hypothesis ($\mathcal{H}$):** I is the CEO of Disney.
- **Hypothesis ($\mathcal{H}$):** Chatter is the CEO of Disney.
- **Hypothesis ($\mathcal{H}$):** Disney is the CEO of Disney.
- **Hypothesis ($\mathcal{H}$):** NBC is the CEO of Disney.
- **Hypothesis ($\mathcal{H}$):** Fox is the CEO of Disney.

FONDAZIONE BRUNO KESSLER

Artificially generated samples

Baselines trained like before

	RTE3	RTE4	RTE5
#	673	315	418
BL 1	.50	.75	.20
BL 2	.15	.25	.12
SPM	.83	.91	.83

Note: It has been proven that, whatever your algorithm is, you can find a data set on which it outperfoms every competitor, by a margin large enough to ensure publication.

(LaLoudouana and Tarare, Data Set Selection. In: Journal of Machine Learning Gossip. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.2501&rep=rep1&type=pdf`)

Conclusions

- SVM approach gives variable results from corpus to corpus.
- Our approach is robust and precise (confident)
- The plan: Increase recall by "reaching out" more and more

Future work

- Extract information from different patterns
- Solve coreference
- Look for contradiction (e.g. antonyms)
- Extend synonimy with mined entailment patterns (Eyal's work)

Thank you for your time and for this amazing opportunity!