Data Science for Deep Natural Language Understanding

Camilo Thorne

Data and Web Science (DWS) Group
Universität Mannheim, Germany
camilo@informatik.uni-mannheim.de

NEC Labs Heidelberg, October 2016


Semantic
Clinical
Language
Data
Processing
Extraction
Chatterbots
Natural
Systems
Web
Activity
Controlled
Information
Based
Guidelines
Conversational
Reasoning
Recognition
Conformance
Complexity
Relation
Science
Knowledge
Checking
Troubleshooting

Motivation

- Meaning is a key dimension of text analytics
- Most semantic analysis approaches rely on lexical semantics and shallow semantics
  - WordNet/Thesauri lexical relationships (synonymy, hypernymy, etc.)
  - Semantic relatedness as per word embeddings/distributional models
  - Shallow semantic analysis/parsing (NER, SRL, etc.)
- However, for some tasks deeper approaches are needed
  - entity linking (linking terms to LOD ontologies)
  - model extraction (extracting consistent symbolic models from text)
  - controlled natural languages (managing ontologies using, e.g., English)

Motivation

- Meaning is a key dimension of text analytics
- Most semantic analysis approaches rely on lexical semantics and shallow semantics
  - WordNet/Thesauri lexical relationships (synonymy, hypernymy, etc.)
  - Semantic relatedness as per word embeddings/distributional models
  - Shallow semantic analysis/parsing (NER, SRL, etc.)
- However, for some tasks deeper approaches are needed
  - entity linking (linking terms to LOD ontologies)
  - model extraction (extracting consistent symbolic models from text)
  - controlled natural languages (managing ontologies using, e.g., English)

Research Goals

Combine knowledge-based and data driven approaches to semantic analysis on specialized domains

Outline

(Clinical) Entity Linking [Thorne et al., 2016]

Low dose **pramipexole** is neuroprotective in
the MPTP mouse model of **Parkinson's disease** (*)

Problems:

1. identify entities (nouns, noun phrases) within an text;
2. identify or resolve the meaning of such entities within such text by linking them to a sense repository
3. resolve meaning of both **domain-specific** and generic terms

(Clinical) Entity Linking [Thorne et al., 2016]

Low dose **pramipexole** is neuroprotective in
the MPTP mouse model of **Parkinson's disease** (*)

Problems:

1. identify entities (nouns, noun phrases) within an text;
2. identify or resolve the meaning of such entities within such text by linking them to a sense repository
3. resolve meaning of both **domain-specific** and generic terms

Question

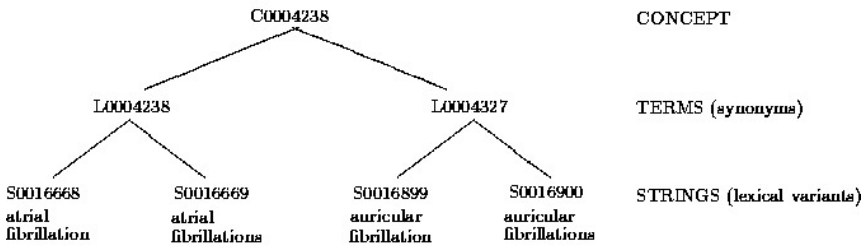
Are there annotation services capable of both?

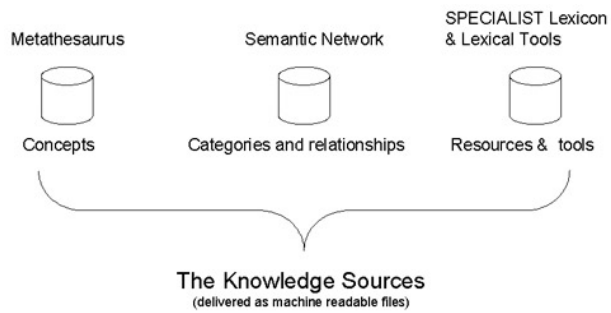
SemRep Corpus [Kilicoglu et al., 2011]

- Experiments ran over the SemRep corpus
- Small annotated clinical corpus
  - 428 clinical excerpts (MedLine/PubMed)
  - $13,948$ word tokens
  - 856 UMLS-annotated clinical terms
- For each sentence, two noun phrases annotated with their corresponding UMLS CUI by clinicians
- 606 terms can be associated to a corresponding DBpedia URI
- Example (*) taken from SemRep

UMLS [Aronson and Lang, 2010]

- Clinical sense repository: Unified Medical Language System Metathesaurus
- Clinical thesaurus comprising:
  1. > 1 million biomedical concepts ($=$ "synsets")
  2. > 5 million concept names ($=$ "senses")
  3. > 100 incorporated controlled vocabularies and classification systems (SNOMED CT, MeSH, SNOMED CT, DSM-IV, LOINC, UK Clinical Terms, RxNorm, Gene Ontology, etc.)
  4. concepts structured as semantic network (linked by clinical relations)





Annotators

MetaMap
[Aronson and Lang, 2010]
clinical domain
sense repository: UMLS
REST service
multilingual
sense: CUI

BabelFly
[Moro et al., 2014]

general domain
sense repository: BabelNet
REST service
multilingual
sense: babelsynset

TagMe
[Ferragina and Scaiella, 2010]
general domain
"sense" repository: Wikipedia
REST service
English/Italian
"sense": Wiki page

WordNet (Lesk)
(custom)

general domain
sense repository: WordNet 3.0
Baseline
English
sense: synset

Annotators

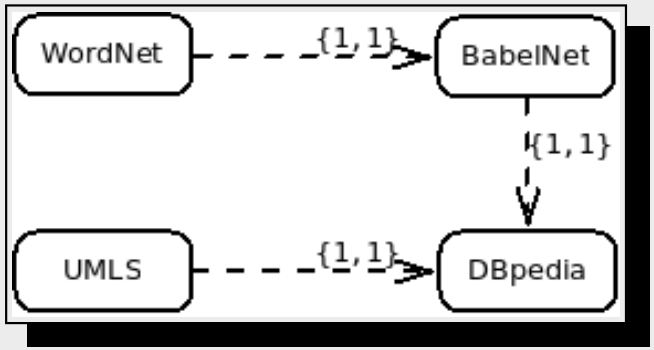
MetaMap	**BabelFly**	**TagMe**	**WordNet (Lesk)**
[Aronson and Lang, 2010]	[Moro et al., 2014]	[Ferragina and Scaiella, 2010]	(custom)
clinical domain	general domain	general domain	general domain
sense repository: UMLS	sense repository: BabelNet	"sense" repository: Wikipedia	sense repository: WordNet 3.0
REST service	REST service	REST service	Baseline
multilingual	multilingual	English/Italian	English
sense: CUI	sense: babelsynset	"sense": Wiki page	sense: synset

Alignment

UMLS can be mapped to DBpedia via Medline and the LikedLifeData initiative [Momtchev et al., 2009]

Annotations [Thorne et al., 2016]

Use DBpedia as pivot:

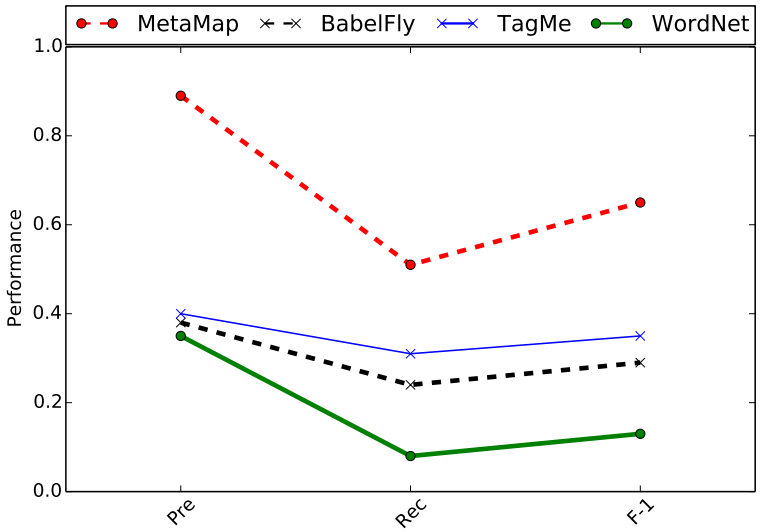


	sense	sense ID	DBpedia URI
Clinical (Gold)	pramipexol Parkinson disease	C0074710 C0030567	*http://dbpedia.org/resource/Pramipexole* *http://dbpedia.org/resource/Parkinson_disease*
MetaMap	pramipexol Parkinson disease	C0074710 C0030567	*http://dbpedia.org/resource/Pramipexole* *http://dbpedia.org/resource/Parkinson_disease*
BabelFly	ATC_code_N04BC05	bn:03124207n	*http://dbpedia.org/resource/Pramipexole*
TagMe	pramipexole Parkinson's disease	*https://goo.gl/twrSVu* *https://goo.gl/Xke6W3*	*http://dbpedia.org/resource/Pramipexole* *http://dbpedia.org/resource/Parkinson's_disease*

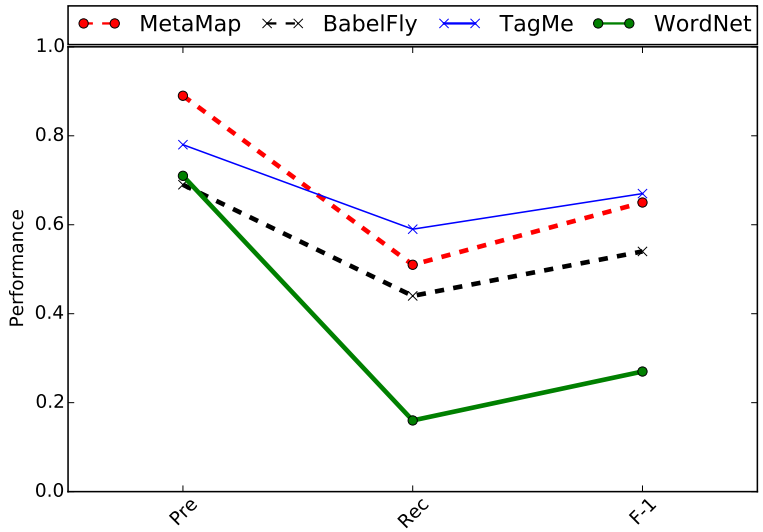
annotations for example (*)

Cross-Evaluation [Thorne et al., 2016]

$$Pre = \frac{\#\text{corrent senses}}{\#\text{returned senses}} \qquad Rec = \frac{\#\text{corrent senses}}{\#\text{corpus senses}} \qquad F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}$$



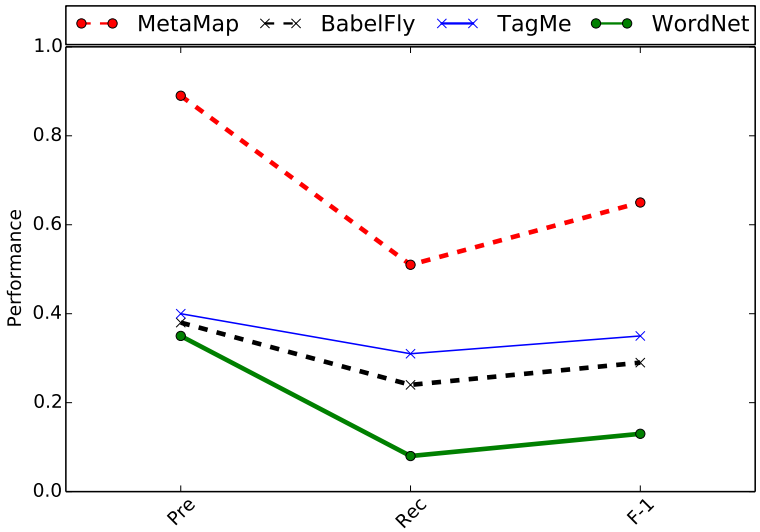
(unresolved URIs)



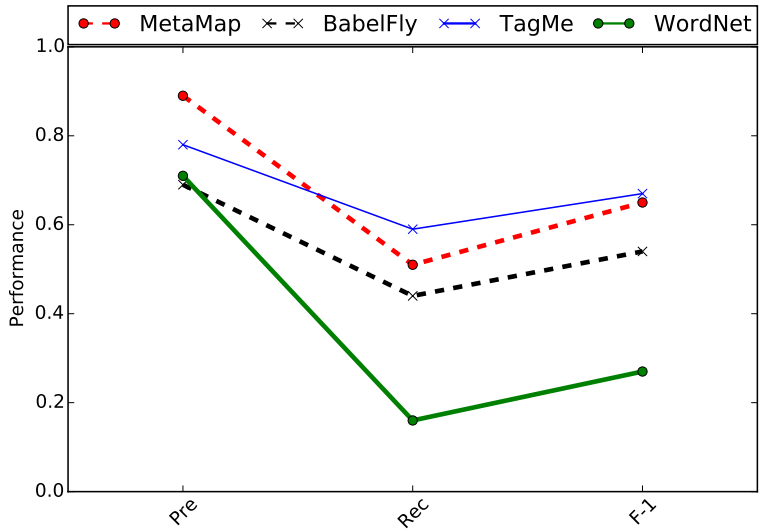
(resolved URIs)

Cross-Evaluation [Thorne et al., 2016]

$$Pre = \frac{\#\text{corrent senses}}{\#\text{returned senses}} \qquad Rec = \frac{\#\text{corrent senses}}{\#\text{corpus senses}} \qquad F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}$$



(unresolved URIs)

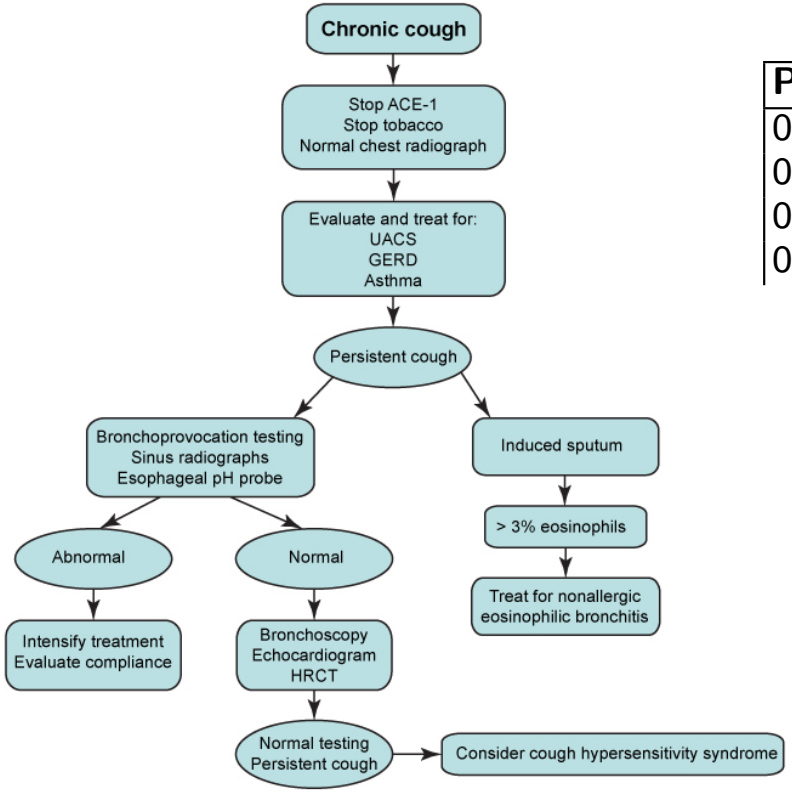


(resolved URIs)

Conclusion

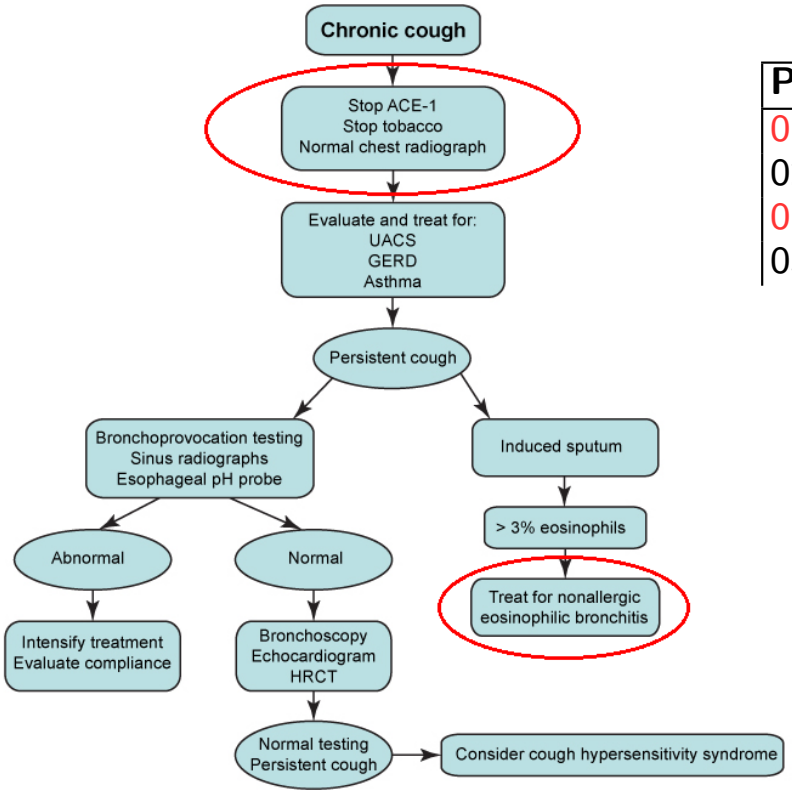
When URIs are resolved via `same as`, generic EL systems such as TagMe and BabelNet match domain-specific annotators like MetaMap

Clinical Guidelines



Pat. ID	Time	Event Type	Loc.	Resource	Personnel
0125	14h00:15	chest-radiograph	A205	radio	A.Hoffer (r.)
0225	15h30:00	chest-radiograph	A205	radio	A.Hoffer (r.)
0125	17h00:10	bronchitis med.	E102	tetracycline	E.Kim (n.)
0333	18h23:05	bronchoscopy	E100	probe	U.Kahn (d.)
⋮	⋮	⋮	⋮	⋮	⋮

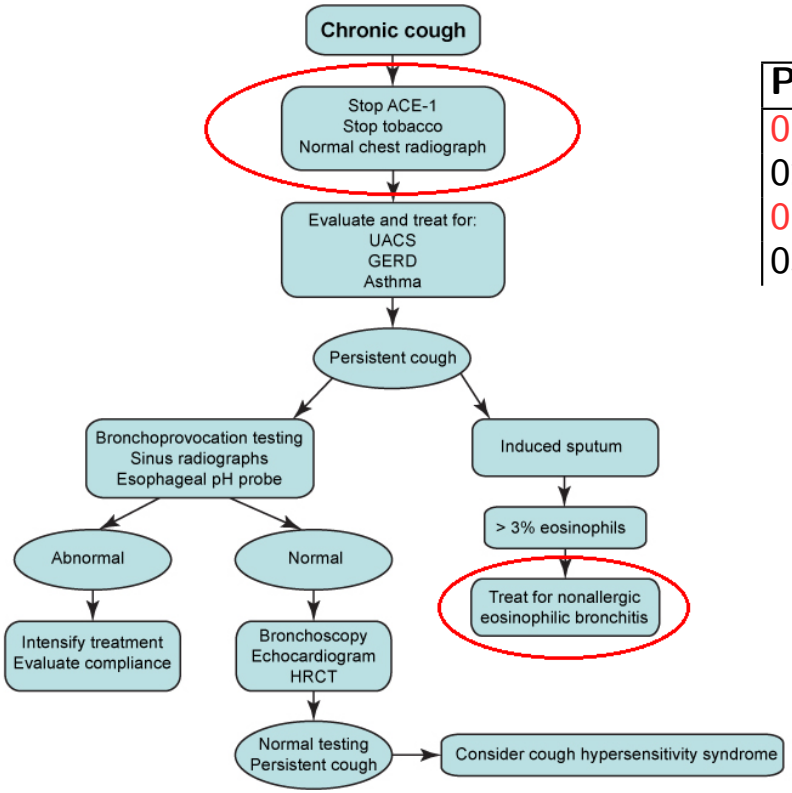
Clinical Guidelines



Pat. ID	Time	Event Type	Loc.	Resource	Personnel
0125	14h00:15	chest-radiograph	A205	radio	A.Hoffer (r.)
0225	15h30:00	chest-radiograph	A205	radio	A.Hoffer (r.)
0125	17h00:10	bronchitis med.	E102	tetracycline	E.Kim (n.)
0333	18h23:05	bronchoscopy	E100	probe	U.Kahn (d.)
⋮	⋮	⋮	⋮	⋮	⋮

violation!!!

Clinical Guidelines



Pat. ID	Time	Event Type	Loc.	Resource	Personnel
0125	14h00:15	chest-radiograph	A205	radio	A.Hoffer (r.)
0225	15h30:00	chest-radiograph	A205	radio	A.Hoffer (r.)
0125	17h00:10	bronchitis med.	E102	tetracycline	E.Kim (n.)
0333	18h23:05	bronchoscopy	E100	probe	U.Kahn (d.)
⋮	⋮	⋮	⋮	⋮	⋮

violation!!!

Question

Use (clinical) NLP annotation to extract CIGs [Kaiser et al., 2007]?

Recognizing Activities and Relations

Definition (Recognition, Abacha and Zweigenbaum [2011])

In the clinical entity and relation recognition task we determine the entities $\bar{t}^*$ s.t.

$$\bar{t}^* = \arg\max_{\bar{t}} \mu(\rho(\bar{\alpha}, \bar{t}))$$

1. t is either a class label c or a relation label r
2. α is a noun phrase (**NP**) or entity
3. $\mu(\cdot)$ is a classifier (e.g., logistic regression)
4. $\rho(\cdot, \cdot)$ is a feature extraction function

Recognizing Activities and Relations

Definition (Recognition, Abacha and Zweigenbaum [2011])

In the clinical entity and relation recognition task we determine the entities $\bar{t}^*$ s.t.

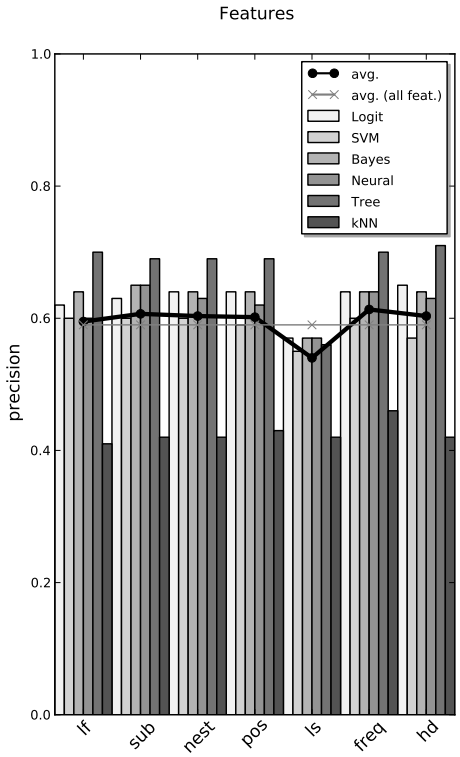
$$\bar{t}^* = \arg\max_{\bar{t}} \mu(\rho(\bar{\alpha}, \bar{t}))$$

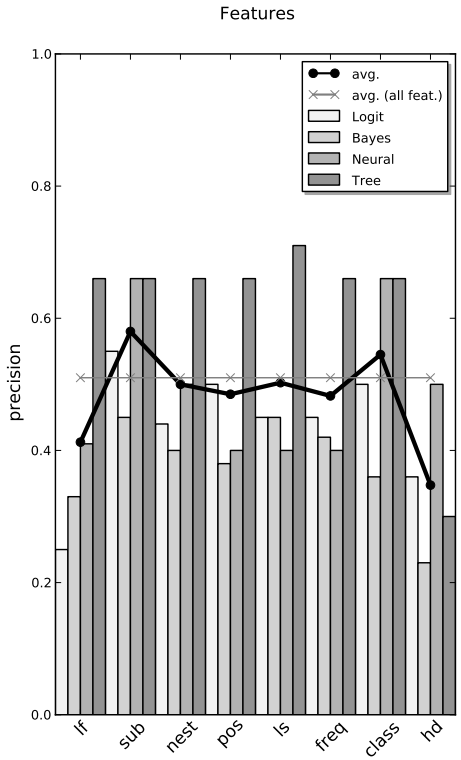
1. t is either a class label c or a relation label r
2. α is a noun phrase (**NP**) or entity
3. $\mu(\cdot)$ is a classifier (e.g., logistic regression)
4. $\rho(\cdot, \cdot)$ is a feature extraction function

We study this task w.r.t. the label sets:

$$\underbrace{\{\text{activity}, \text{resource}, \text{actor}, \text{other}\}}_{\text{entity type label } c} \quad \underbrace{\{\text{temporal}, \text{causal}, \text{other}\}}_{\text{relation label } r}$$

Recognition Experiments [Thorne et al., 2013b,c]





(1) activity precision

χ^2	p-value
408242025.00	0.0

(2) relation precision

χ^2	p-value
18550249.00	0.0

Null hypothesis H_0: uniform distribution

Controlled English [Thorne et al., 2013a]

- English interfaces to OWL ontologies

Every Italian loves some pasta	$\mathit{Italian} \sqsubseteq \exists \mathit{loves}.\mathit{Pasta}$
Silvio is Italian	$\mathit{Italian}(\text{Silvio})$
$\therefore$ Silvio Loves some pasta	$\therefore \exists \mathit{loves}.\mathit{Pasta}(\text{Silvio})$

Controlled English [Thorne et al., 2013a]

- English interfaces to OWL ontologies

Every Italian loves some pasta	$\mathit{Italian} \sqsubseteq \exists \mathit{loves}.\mathit{Pasta}$
Silvio is Italian	$\mathit{Italian}(\text{Silvio})$
$\therefore$ Silvio Loves some pasta	$\therefore \exists \mathit{loves}.\mathit{Pasta}(\text{Silvio})$

- Require deep semantic analysis of English

Controlled English [Thorne et al., 2013a]

- English interfaces to OWL ontologies

Every Italian loves some pasta	$\mathit{Italian} \sqsubseteq \exists \mathit{loves}.\mathit{Pasta}$
Silvio is Italian	$\mathit{Italian}(\text{Silvio})$
$\therefore$ Silvio Loves some pasta	$\therefore \exists \mathit{loves}.\mathit{Pasta}(\text{Silvio})$

- Require deep semantic analysis of English
- But, natural language is ambigous $\Rightarrow$ blows up processing

Controlled English [Thorne et al., 2013a]

- English interfaces to OWL ontologies

Every Italian loves some pasta	$\mathit{Italian} \sqsubseteq \exists \mathit{loves}.\mathit{Pasta}$
Silvio is Italian	$\mathit{Italian}(\text{Silvio})$
$\therefore$ Silvio Loves some pasta	$\therefore \exists \mathit{loves}.\mathit{Pasta}(\text{Silvio})$

- Require deep semantic analysis of English
- But, natural language is ambigous $\Rightarrow$ blows up processing

Solution

Tackle the ambiguity problem with controlled fragments of English

Controlled Fragments [Pratt-Hartmann and Third, 2006]

Controlled Fragment	Constructs	Complexity
COP	Copula, common and proper nouns, negation, universal, existential quantifiers	**PTime**
COP+Rel	COP plus relative pronouns	**NP**-complete
COP+TV	COP plus transitive verbs	**PTime**
COP+TV+DTV	COP+TV plus ditransitive verbs	**PTime**
COP+Rel+TV	COP+Rel plus transitive verbs	**ExPTime**-complete
COP+Rel+TV+DTV	COP+Rel+TV plus ditransitive verbs	**ExPTime**-complete
COP+Rel+TV+RA	COP+Rel+TV plus anaphoric pronouns (e.g., he, him, it, herself) of bounded scope	**ExPTime**-complete
COP+Rel+TV+GA	COP+Rel+TV plus unbounded anaphoric pronouns	undecidable
COP+Rel+TV+DTV+RA	COP+Rel+TV+DTV plus bounded anaphoric pronouns	undecidable

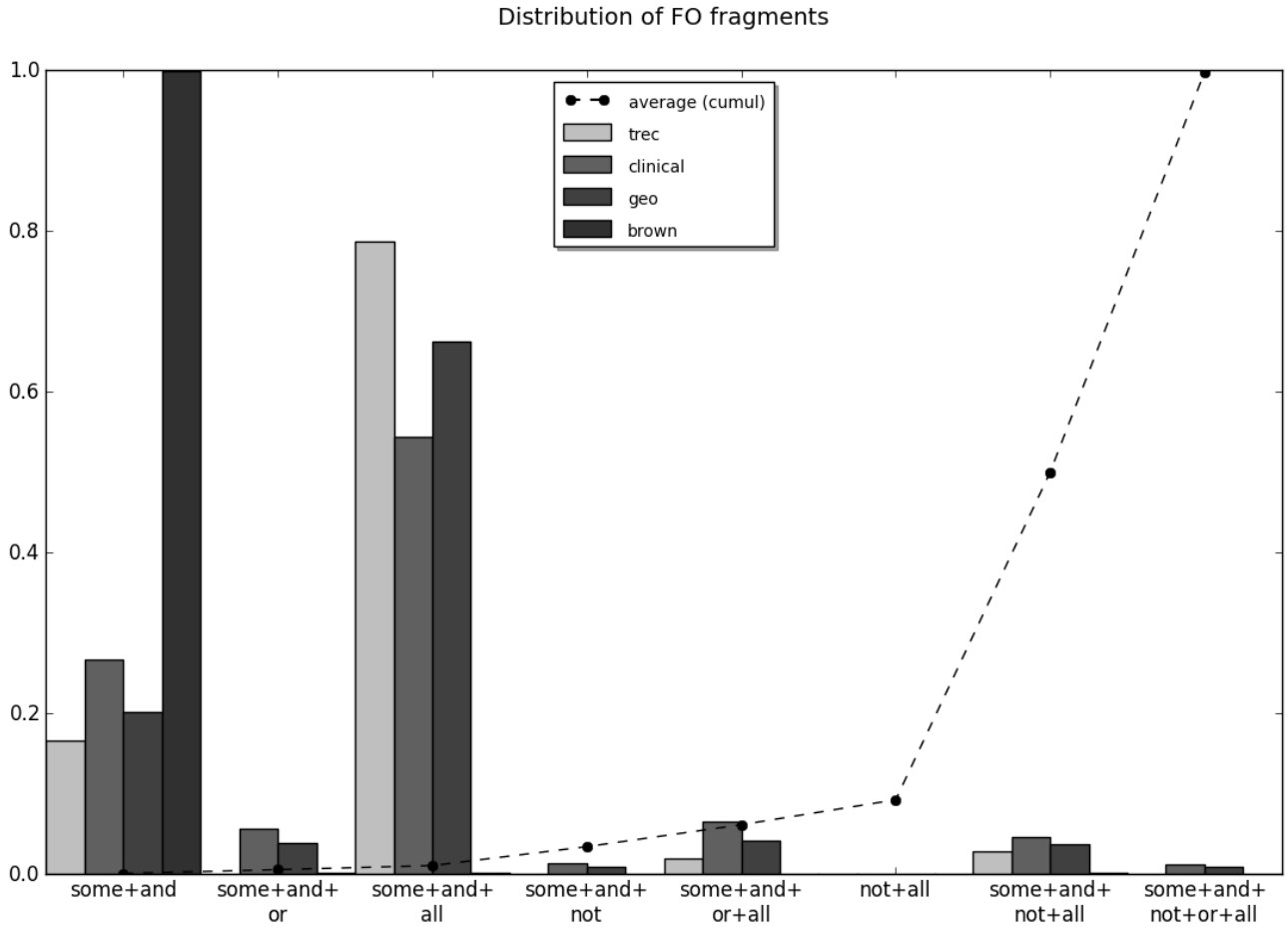
Controlled Fragments [Pratt-Hartmann and Third, 2006]

Controlled Fragment	Constructs	Complexity
COP	Copula, common and proper nouns, negation, universal, existential quantifiers	**PTime**
COP+Rel	COP plus relative pronouns	**NP**-complete
COP+TV	COP plus transitive verbs	**PTime**
COP+TV+DTV	COP+TV plus ditransitive verbs	**PTime**
COP+Rel+TV	COP+Rel plus transitive verbs	**ExPTime**-complete
COP+Rel+TV+DTV	COP+Rel+TV plus ditransitive verbs	**ExPTime**-complete
COP+Rel+TV+RA	COP+Rel+TV plus anaphoric pronouns (e.g., he, him, it, herself) of bounded scope	**ExPTime**-complete
COP+Rel+TV+GA	COP+Rel+TV plus unbounded anaphoric pronouns	undecidable
COP+Rel+TV+DTV+RA	COP+Rel+TV+DTV plus bounded anaphoric pronouns	undecidable

Question

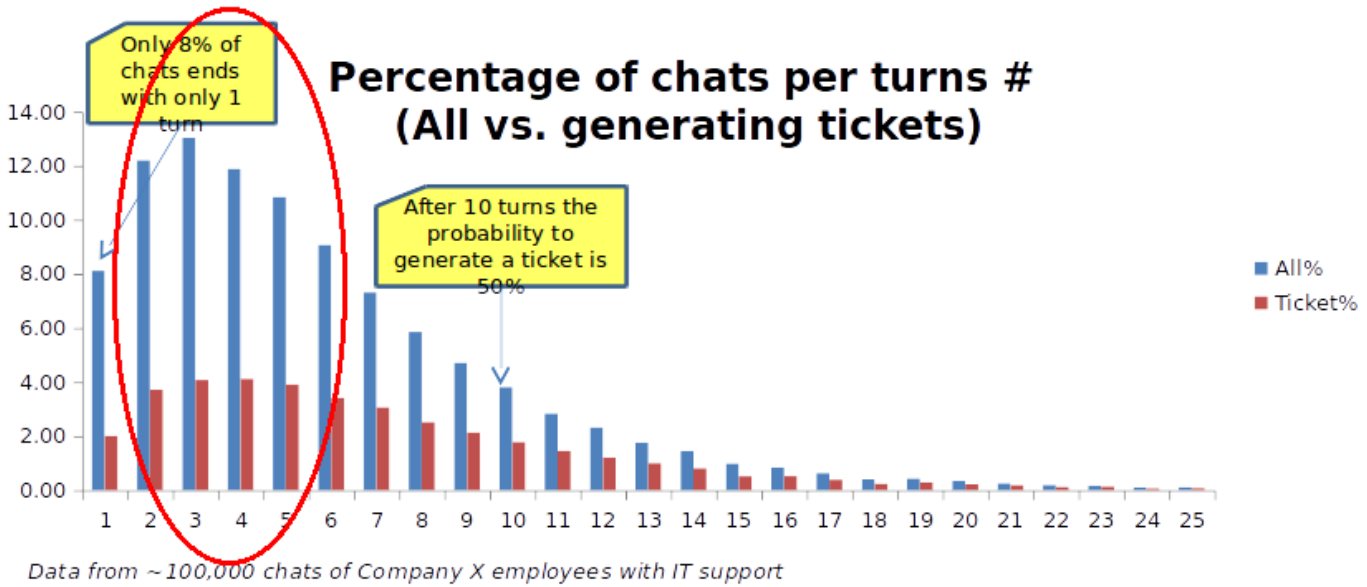
How often do we use the cheapest constructs?

Controlled Language Distribution [Thorne et al., 2013a]



Chatbots for IT Troubleshooting (IBM Italia)

- Developed a conversational agent proof-of-concept for IT troubleshooting chats
- Dialog management driven by knowledge-based reasoning and centering
- Extended the Italian IBM Watson® NLP stack to process text-based conversation




Thank You!

References I

A. B. Abacha and P. Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of the BioNLP 2011 Workshop*, 2011.

A. R. Aronson and F.-M. Lang. And overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3): 229–236, 2010.

P. Ferragina and U. Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 2010.

K. Kaiser, C. Akkaya, and S. Miksch. How can information extraction ease formalizing of treatment processes in clinical practice guidelines? *Artificial Intelligence in Medicine*, 39(2):151–163, 2007.

References II

H. Kilicoglu, G. Rosenblat, M. Fiszman, and T. C. Rindfleisch. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, 12 (486), 2011.

V. Momtchev, D. Peychev, T. Primov, and G. Georgiev. Expanding the pathway and interaction knowledge in linked life data. *Proceedings of 2009 International Semantic Web Challenge*, 2009.

A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2: 231–244, 2014.

I. Pratt-Hartmann and A. Third. More fragments of language. *Notre Dame Journal of Formal Logic*, 47(2):151–177, 2006.

C. Thorne, R. Bernardi, and D. Calvanese. Designing efficient controlled languages for ontologies. In H. Bunt, editor, *Computing Meaning, Volume 4*, pages 149–173. Springer, 2013a.

References III

C. Thorne, E. Cardillo, M. Montali, C. Eccher, and D. Calvanese. Automated activity recognition in clinical documents. In *Proceedings of the 6th International Joint Conference in Natural Language Processing (IJCNLP 2013)*, 2013b.

C. Thorne, E. Cardillo, M. Montali, C. Eccher, and D. Calvanese. Process fragment recognition in clinical documents. In *Proceedings of the 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013)*, 2013c.

C. Thorne, S. Faralli, and H. Stuckenschmidt. Cross-evaluation of entity linking and disambiguation systems for clinical text annotation. In *Proceedings of the 12th International Conference on Semantic Systems (SEMANTiCS2016)*, 2016.