~~Overview: Statistics-Focused Approaches to NLP~~

Christopher Manning

Stanford University

http://nlp.stanford.edu/~manning/

NLM workshop on NLP: State of the Art, Future Directions, 2012

Language Understanding-Focused Approaches to NLP

Christopher Manning

Stanford University

http://nlp.stanford.edu/~manning/

NLM workshop on NLP: State of the Art, Future Directions, 2012

Okay, there has been a lot of probabilistic models and machine learning lately…

Commencement 2010

The idea of language as a discrete, logical system

"Ordinary mathematical techniques fall mostly into two classes, the continuous (e.g., the infinitesimal calculus) and the discrete or discontinuous (e.g., finite group theory). Now it will turn out that the mathematics called 'linguistics' belongs to the second class. It does not even make any compromise with continuity as statistics does, or infinite-group theory. Linguistics is a quantum mechanics in the most extreme sense. All continuities, all possibilities of infinitesimal gradation, are shoved outside of linguistics in one direction or the other."

Joos (1950: 701–702)

Categorical linguistic theories claim too much

They assert a hard categorical boundary of grammaticality, where really there is a fuzzy edge,

determined by many conflicting constraints and issues of conventionality vs. human creativity

“All grammars leak.”

(Sapir 1921: 38)

Categorical linguistic theories explain too little

They say nothing at all about the soft constraints which explain how people choose to express things

Something that computational linguists – and anyone else dealing with real language use – usually want to know about

“Statistical considerations are essential to an understanding of the operation and development of languages”

(Lyons 1968: 98)

Statistical Language Universals

The case of coordination

Generative linguistics proposed coordination as a categorically *internally* constrained syntactic relation:

- *a boy and his dog* **vs.** **a boy and happy*
- *Conjoin likes* (Chomsky 1965): X → X and X

As a categorical claim, the rule is empirically **false**:

- *52 years* ***old and a*** *27-year Reuters* ***veteran***
- *not* ***cruddy****,* ***but*** *not* ***a dress*** *either*

Modern formal linguistics tries to save the day with a weaker theory of *extrinsic* constraints (Ingria 1990, Sag 2002):

- Each conjunct satisfies external case, category constraints

But, *conjoin likes* is **true** as a **statistical** claim

- Interpreting it statistically *increases* its explanatory power
- The statistical claim *cannot* be captured by an external constraint

Probabilities in Linguistics

There is now active work exploring such ideas in many areas of linguistics

Syntax

Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English

Semantics

Semantic factors in the choice between ethnic adjectives and PP counterparts: Quantitative evidence

Phonology

Stochastic phonological knowledge: the case of Hungarian vowel harmony

How I got into Statistical NLP

Writing an English grammar for a parser

(LFG for the Xerox PARC GWW in the early 1990s)

The problem of coverage vs. ambiguity

How could you work out the right structure … and later interpretation for a particular sentence?

Language understanding

Language understanding is a process of flexibly reasoning under uncertainty

Human languages are ambiguous, people speak ambiguously, the interpreter has incomplete information, and has to make guesses … based on both the language used and shared context and knowledge

Probabilities help you make good guesses

Linguistic representations

Using probabilities doesn’t mean that we don’t need good linguistic representations

You can – and should – put probabilistic models over complex linguistic representations

Actually, the work I’m best-known for is mainly about linguistic representations

NLP: THE STATE OF THE ART

What have we learned?

Crude text statistics calculated over large amounts of text can often get you a long way …

Protein-protein interaction?

- Scan a lot of text looking for sentences mentioning 2 proteins
- Assume they interact
- Accumulate counts
- Filter resulting list by frequency or mutual information

… a distance undreamed of in early NLP work in the 1970s and 80s

What have we learned?

These kinds of methods ***are*** the appropriate baseline for all your language processing needs

If using them is sufficient for your needs, why do more?

If supposedly deeper and cleverer methods don't perform as well as these methods, you should definitely put them back on the shelf

If you need better performance than this (more **precision** *and/or* **recall**), then you need models with real NLP in them

Then I'm your man!

Finding things with names in text

Biology has *lots* of things with names:

human granulocyte-macrophage CSF	IL-2
nuclear factor-kappa B	for
alpha-naphthyl acetate esterase	eyeless

We call finding them “named entity recognition”

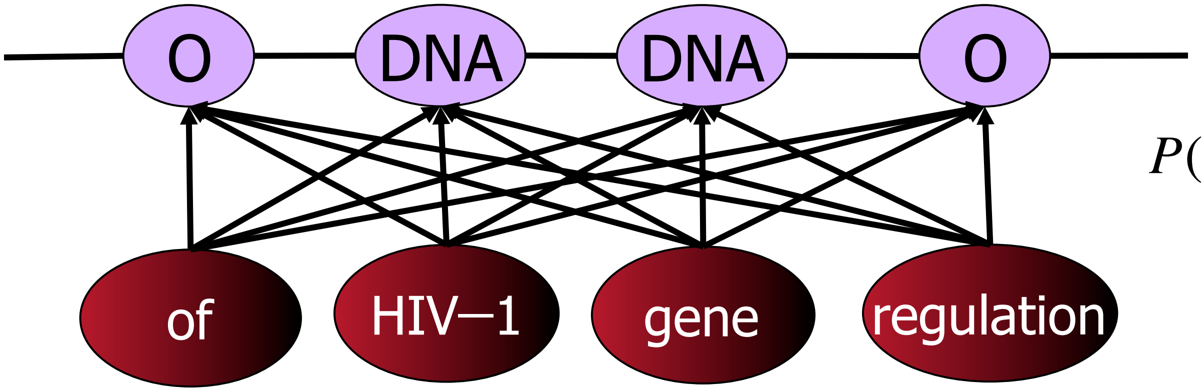
- We probably shouldn’t, but the name has stuck…

It’s a new invention!

- Machine Translation … 1950s
- Syntactic parsing … 1960s
- Named entity recognition … 1996

Probabilistic NER: Conditional Random Fields

[Lafferty, McCallum, & Pereira 2001]



$$P(c \mid w) = \frac{\exp[\sum_{j=1}^{m} f_j(c,w)\lambda_j]}{\sum_{c'} \exp[\sum_{j=1}^{m} f_j(c',w)\lambda_j]}$$

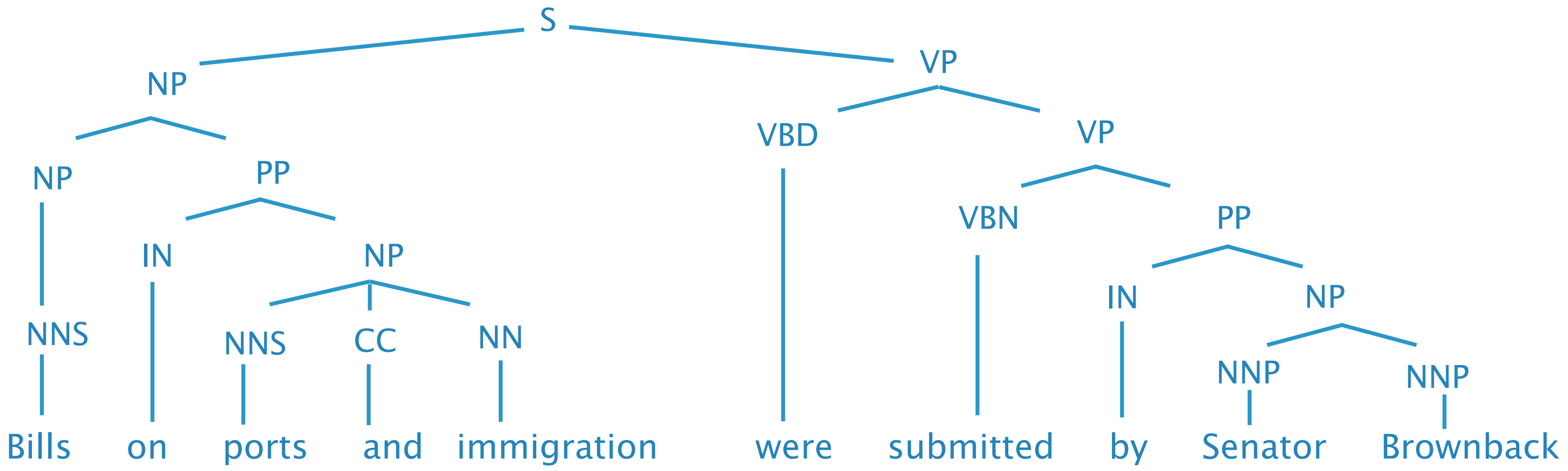
- CRFs give useful and good, but imperfect, systems for NER
  - Newswire NLP: 90+% accuracy; biomedical NLP: mid-80's accuracy
- Name lists take you some distance but probabilistic models:
  - Disambiguate entity status and type based on context
  - Generalize well to new entities based on context and name

Statistical parsing

- One of the big successes of 1990s statistical NLP was the development of statistical parsers
- These are trained from hand-parsed sentences ("treebanks"), and know statistics about phrase structure and word relationships, and use them to assign the most likely structure to a new sentence
- They will return a sentence parse for *any* sequence of words
- It will usually be *mostly right*
- There are many opportunities for exploiting this richer level of analysis, which have only been partly realized.

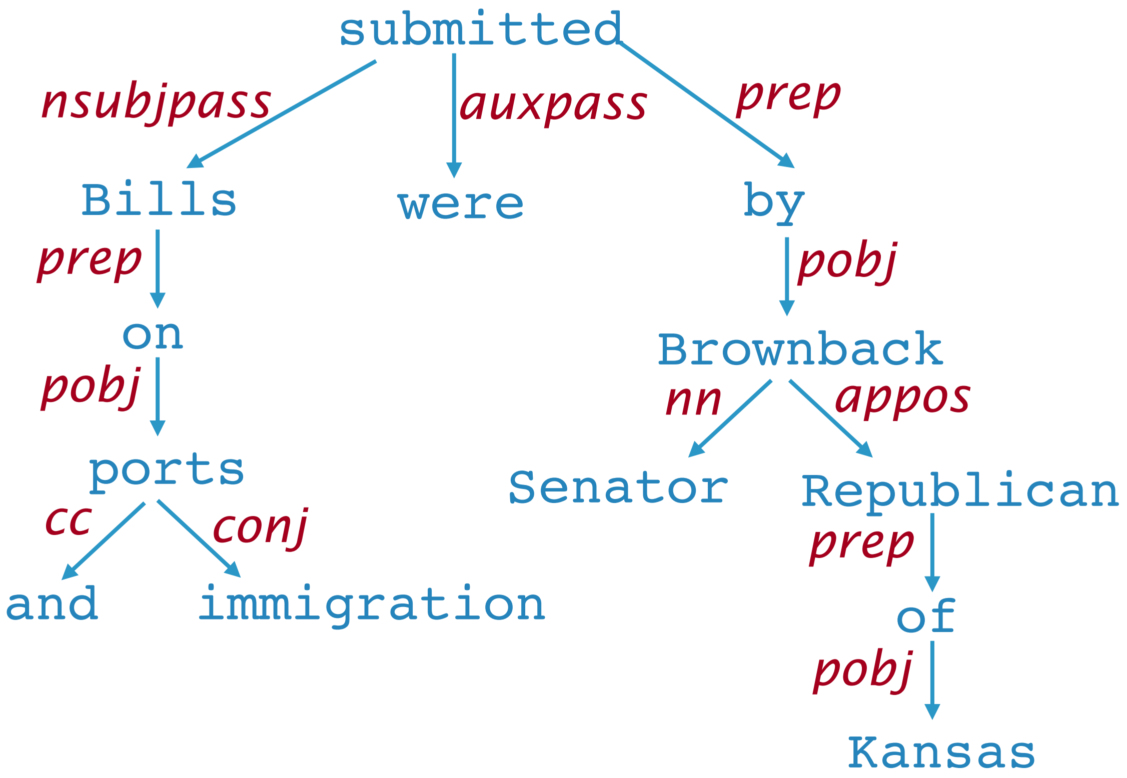
Phrase structure Parsing

- Phrase structure representations have dominated American linguistics since the 1930s
- They focus on showing words that go together to form natural groups (constituents) that behave alike



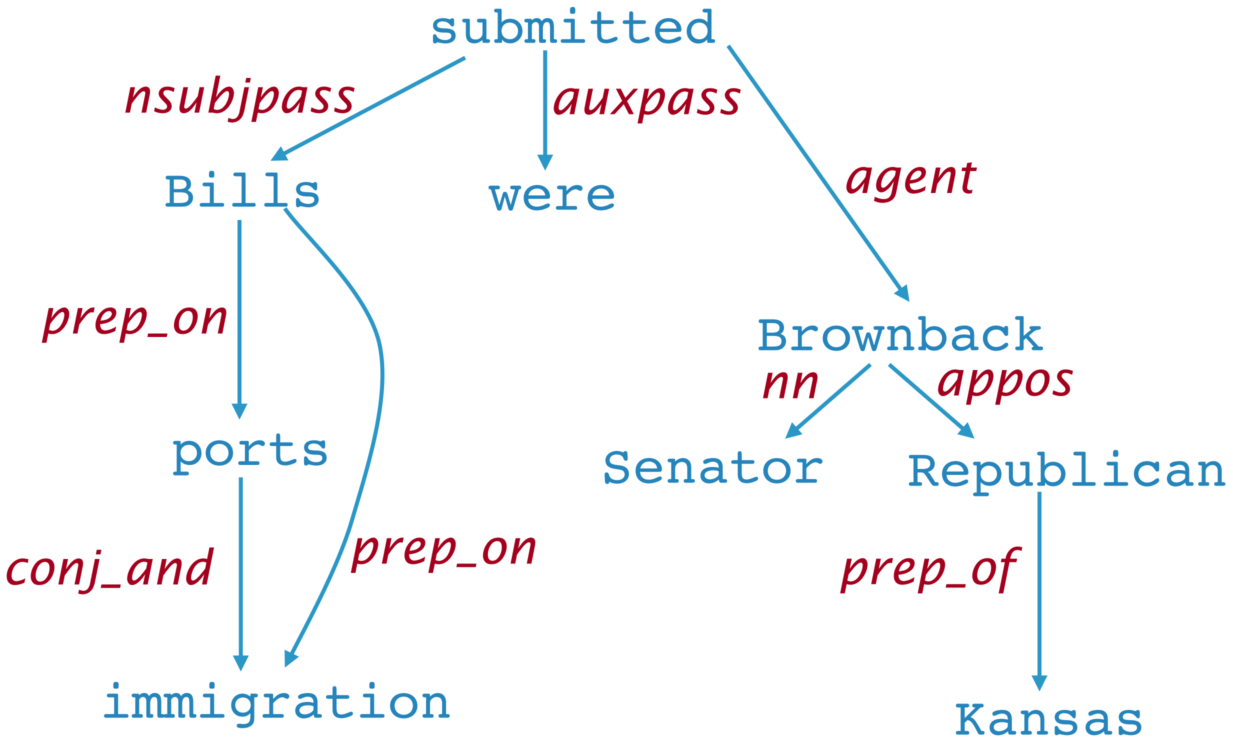
Dependency parsing

- A dependency parse shows which words in a sentence modify other words
- The key notions are governors with dependents



Stanford Dependencies

- SD is a particular dependency representation designed for easy extraction of meaning relationships [de Marneffe & Manning, 2008]
  - It's basic form in the last slide has each word as is
  - A "collapsed" form focuses on relations between content words



Relation extraction

Learning predicates consisting of a relation name and one or more arguments:

> "We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."

- interacts(CBF-A, CBF-C)
- associates(CBF-B, CBF-A-CBF-C complex)

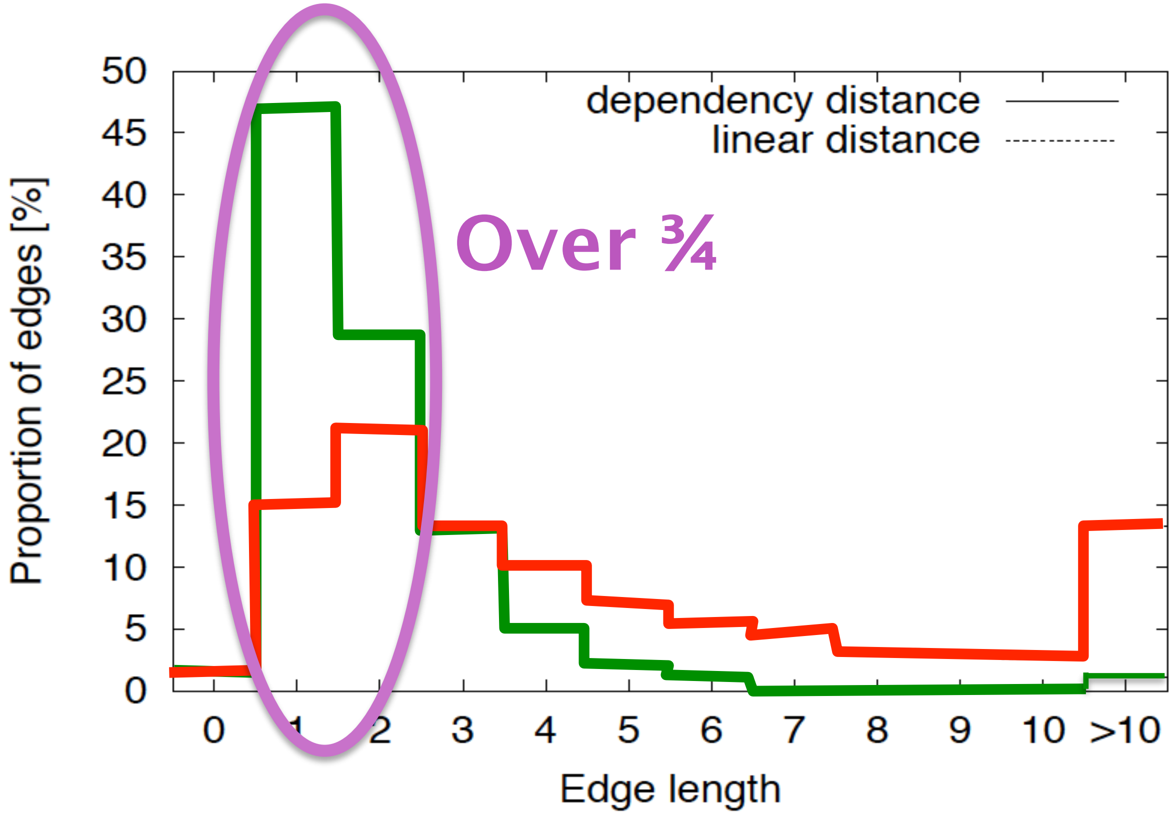
Relation extractors can be trained from supervised data

- often exploiting representations like parsers

Accuracies are reasonable, but more modest than for NER

Stanford Dependencies as a representation for relation extraction

- Stanford Dependencies favor short paths between related content words, and were widely used in the BioNLP shared tasks

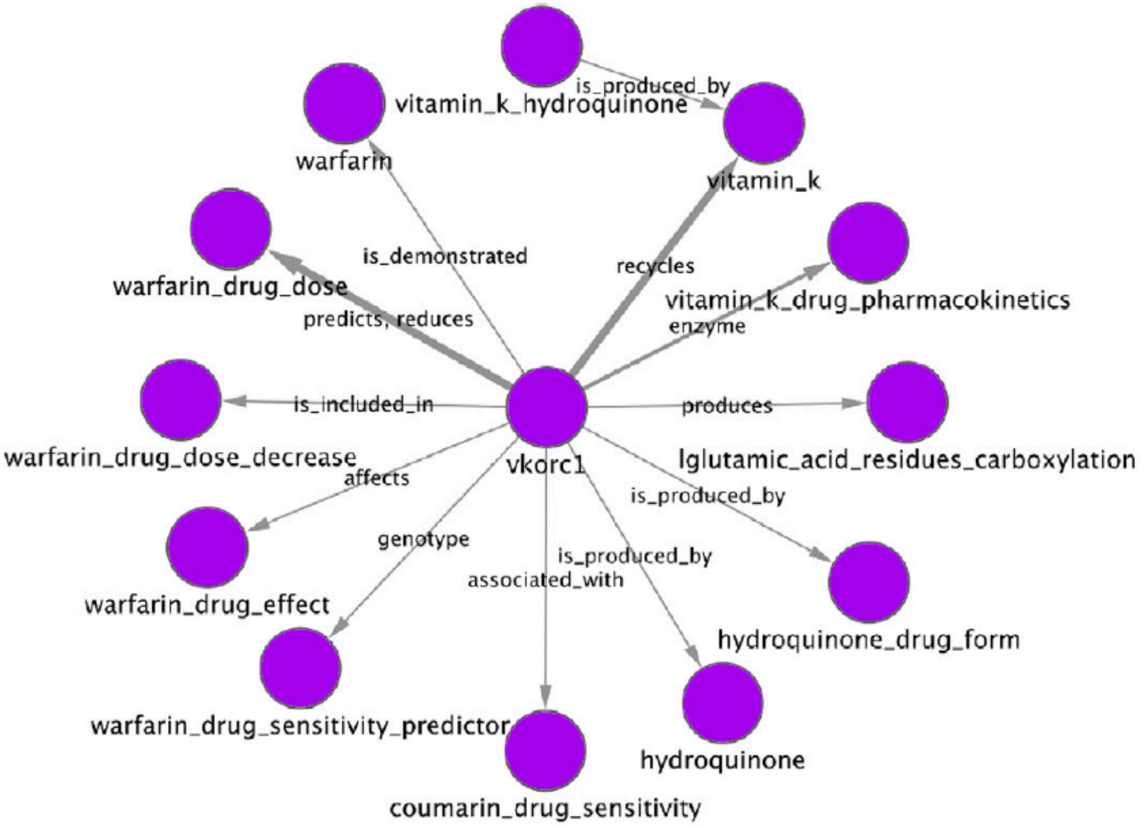


Over ¾

Björne et al. 2009

Dependencies are close to semantic networks

- A. Coulet et al (2010) Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*



Coreference resolution

- The goal is to work out which (noun) phrases refer to the same entities in the world
  - Sarah asked her father to look at her. He appreciated that his eldest daughter wanted to speak frankly.
- ≈ anaphora resolution ≈ pronoun resolution ≈ entity resolution

- Clearly important to text interpretation
- But performance is again still modest
- And it can be hard to get value from it in applications

Stanford CoreNLP

http://nlp.stanford.edu/software/corenlp.shtml

Stanford CoreNLP is our new-ish package that ties together a bunch of our NLP tools

- Tokenization/sentence-splitting/lemmatization
- POS tagging
- Named Entity Recognition
- Parsing
- *and* Coreference Resolution

It has a state-of-the-art coreference system! (CoNLL 2011 winner)

It's completely rule-based

NLP: FUTURE DIRECTIONS

Going beyond the supervised learning paradigm

The NLP successes of the last 15 years were mainly fueled by the supervised learning paradigm

- Spend person months/years annotating text with some classifications/structure suitable for a task
- Train supervised classifier to reproduce these annotations

It was a very successful paradigm

- Sometimes the datasets showed surprising reuse
- E.g., treebanks

It doesn’t remove linguistic knowledge, it just externalizes it

Going beyond the supervised learning paradigm

- But the paradigm has its limits
- It's very costly to undertake each new task
- Or to have systems that work well in different domains
- It means that the training data is always effectively small, despite the fact that we are swimming in big data
- We need to build systems that require less supervision

Relation Extraction from Distant Supervision

[Mintz, et al. ACL 2009; Surdeanu et al. 2011]

- If we had relations marked in texts, we could train a conventional relation extraction system …
- Can we exploit the abundant found information about relations – whether Wikipedia or the Gene Ontology – to be able to bootstrap systems for machine reading?
- Method: use database as “distant supervision” of text
- The challenge is dealing with the “noise” that enters the picture

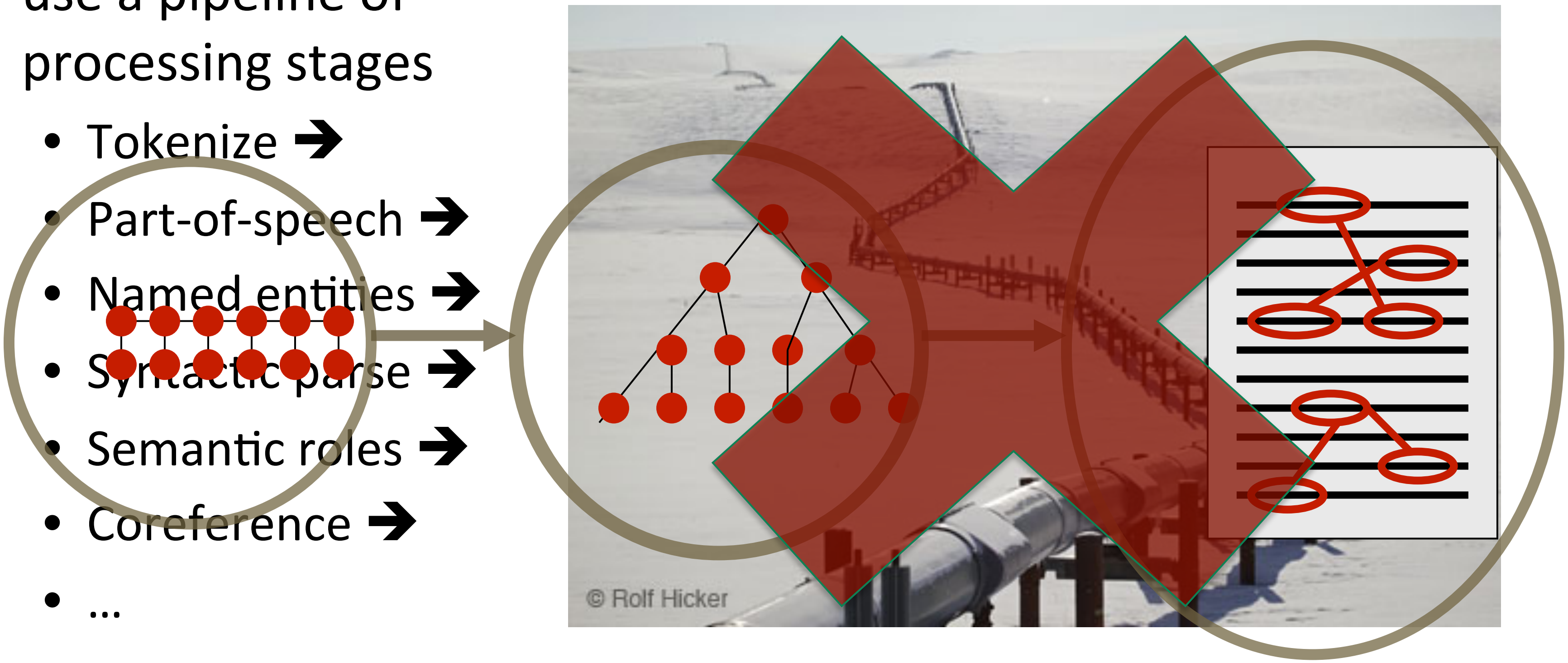
Example Results (using Freebase)

- Precision of extracted facts: about 70%
- New relations learned:

Montmartre **IS-IN** Paris	Fyoder Kamesnky **DIED-IN** Clearwater
Fort Erie **IS-IN** Ontario	Utpon Sinclair **WROTE** Lanny Budd
Vince McMahon **FOUNDED** WWE	Thomas Mellon **HAS-PROFESSION** Judge

How do we design a human language understanding system?

- Most current systems use a pipeline of processing stages
  - Tokenize ➔
  - Part-of-speech ➔
  - Named entities ➔
  - Syntactic parse ➔
  - Semantic roles ➔
  - Coreference ➔
  - …



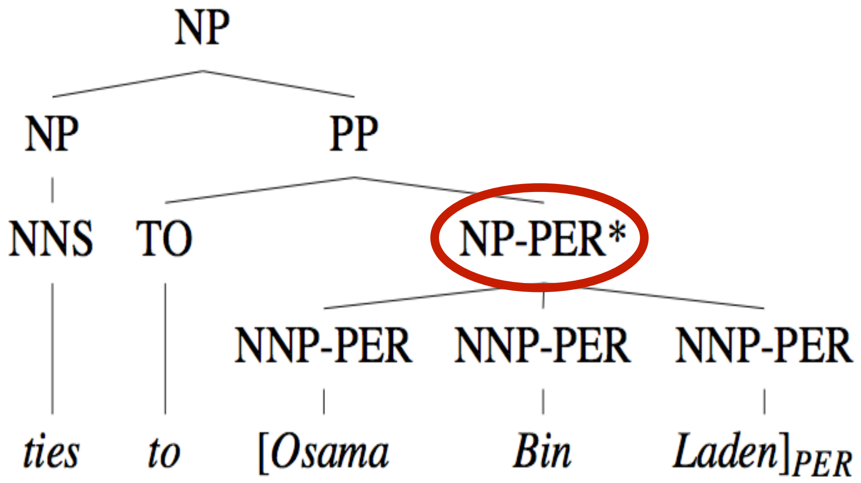
© Rolf Hicker

From Processing Pipelines to Joint Inference

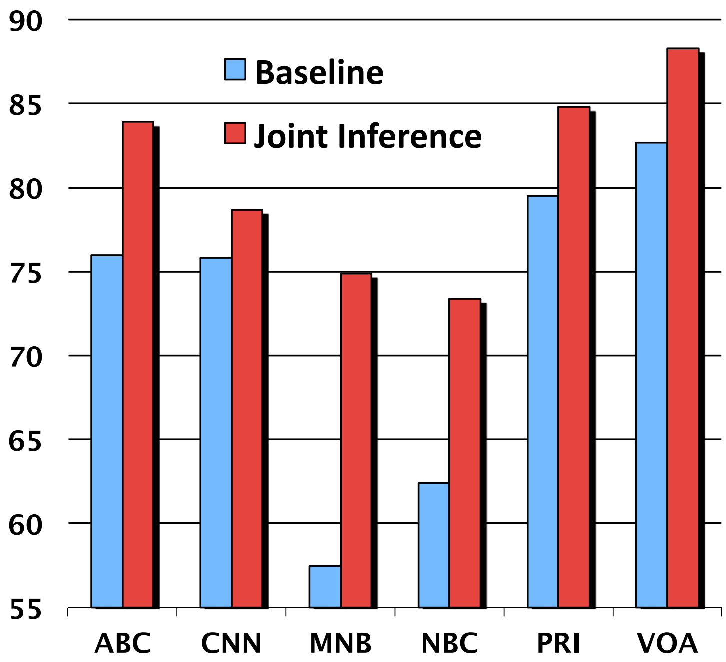
[Finkel & Manning, NAACL 2009, 2010]

- Goal: Joint modeling of the many phases of linguistic analysis
  - Here, parsing and named entities

- Fixed 24% of named entity boundary errors and of incorrect label errors
- 22% improvement in parsing scores



Named Entity Recognition F1-score on OntoNotes (by section)



Towards semantics: How can we understand relationships between pieces of text?

- Can one conclude one piece of text from another?
  - Emphasis is on handling the variability of linguistic expression

- This textual inference technology would enable:
  - Semantic search: *lobbyists attempting to bribe U.S. legislators*

    The A.P. named two more senators who received contributions engineered by lobbyist Jack Abramoff in return for political favors.
  - Question answering: *Who bought J.D. Edwards?*

    Thanks to its recent acquisition of J.D. Edwards, Oracle will soon be able …
  - Clinical report interpretation
  - Concept, paraphrase and contradiction detection

Natural Logic for semantic composition

P	*Jimmy Dean*	*refused to*			*move*	*without*	*blue*	*jeans*
H	*James Dean*		*did*	*n't*	*dance*	*without*		*pants*
edit index	1	2	3	4	5	6	7	8
edit type	SUB	DEL	INS	INS	SUB	MAT	DEL	SUB
lex feats	strsim= 0.67	implic: –/o	cat:aux					hyper
lex entrel	=	|	=					⊏
projec-tivity	↑	↑	↑					↑
atomic entrel	=	|	=	^	⊏	=	⊏	⊏
compo-sition	=	|	|	⊏	⊏	⊏	⊏	⊏

For example:

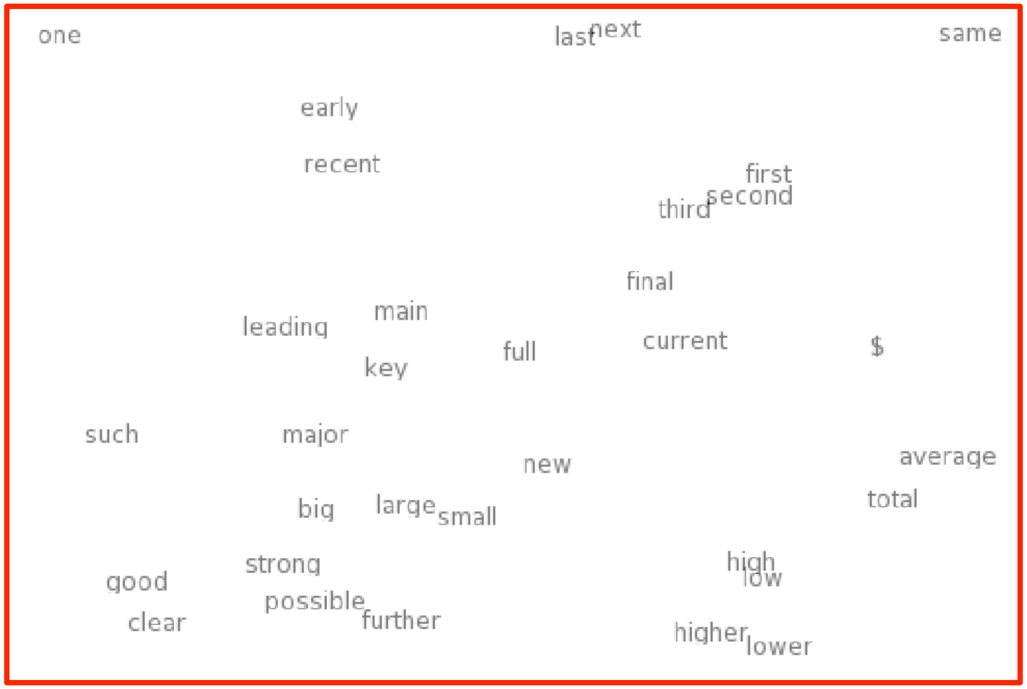
human ^ *nonhuman*

fish | *human* → *fish* < *nonhuman*

Final answer

Distributional learning of meaning

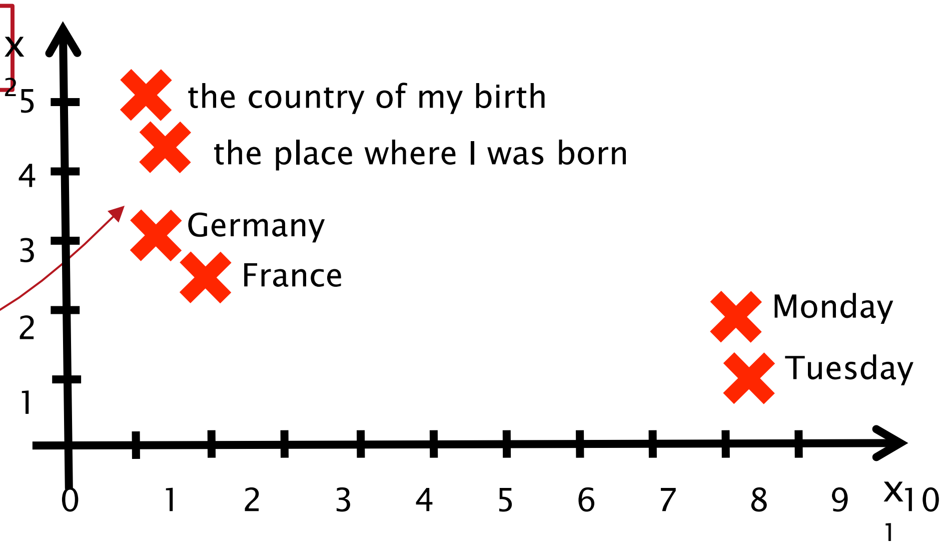
- Distributional learning of word meaning has been very successful – good, operational lexical semantics
- How can we extend this to meaning in language in general?

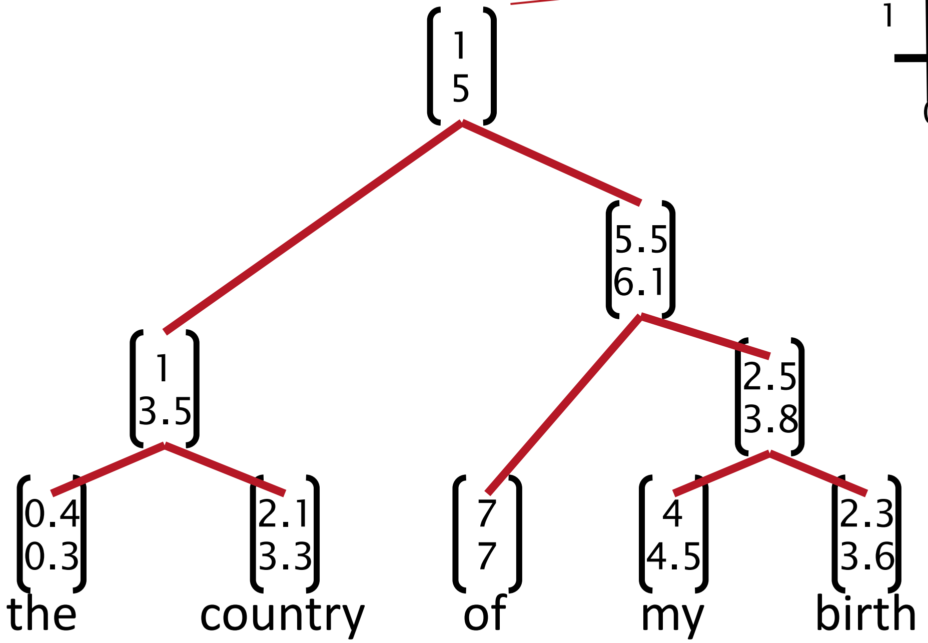


How should we map phrases into a vector space?

Use the principle of compositionality!

The meaning (vector) of a sentence is determined by
(1) the meanings of its words and
(2) the rules that combine them.





Deep Learning algorithms jointly learn compositional vector representations and tree structure.

Language is inherently connected to human communication

> "… the common misconception [is] that language use has primarily to do with words and what they mean.
>
> It doesn't. It has primarily to do with people and what *they* mean."
>
> *Asking questions and influencing answers*
>
> Clark & Schober, 1992

This will typically be the situation when interpreting clinician's reports

➔ **Computational Pragmatics** (de Marneffe, Potts, and Manning 2011)

Veridicality matters

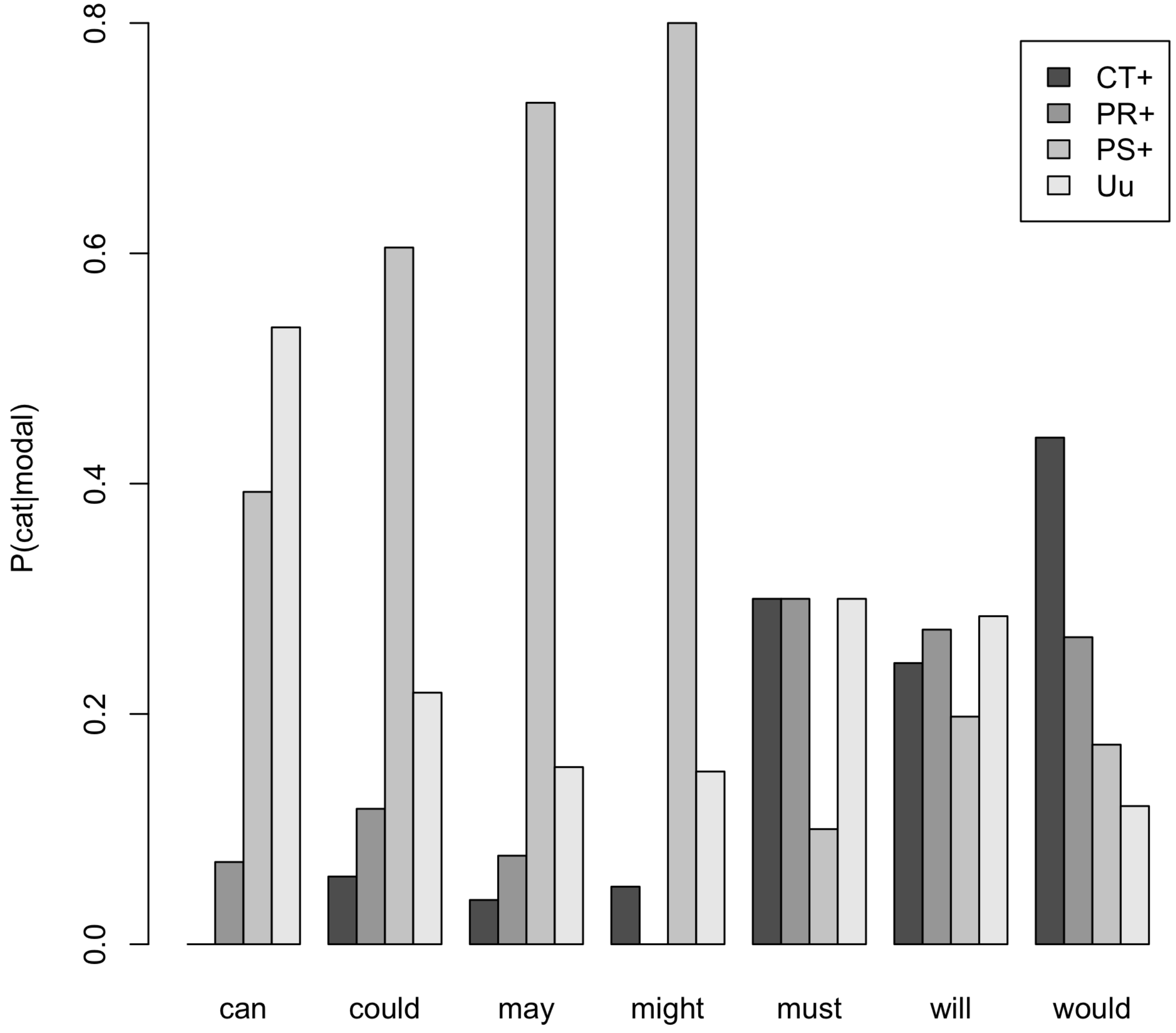
In this study the researchers found no evidence that

the phosphorylation of TRAF2 favors binding to the

CD40 cytoplasmic domain.

→ Does the relation correspond to a factual situation in the real world?

More nuanced veridicality values for modals



Envoi

- Probabilistic models have given us very good tools for analyzing human language sentences
- We can extract participants and their relations with good accuracy
- There is exciting work in text understanding and inference based on these foundations
- This provides a basis for computers to do higher-level tasks that involve knowledge & reasoning
- But much work remains to achieve the language competence of science fiction robots….

LELAND STANFORD JUNIOR UNIVERSITY

DIE LUFT DER FREIHEIT WEHT