Text Analytics

Relationship Extraction

Ulf Leser

Content of this Lecture

- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
  - Protein-Protein-Interactions

Relationship Extraction

- Very often, entities in a sentence are in a certain relationship to each other : Relationship extraction (RE)
  - Price of a product
  - CEO of a company
  - Who bought what?
  - Who talked to whom?
  - Of which band is this song?
  - Which proteins interact with which other proteins?
  - ...
- Usually, RE depends on pre-recognized entities
  - Can be modelled as joint inference problem

Binary versus n-ary RE

Z-100 is an arabinomannan extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv ...

The death toll in an earthquake in south west China is now at least 32, with 467 injuries, state media says."

- [south-west china, death, 32]
- [south-west china, injury, 467]

What to Extract? Types of RE Problems

- Only the entities that have a certain relation
  - Output: Tuples (mostly pairs) of entities
  - Usually implicitly defined through training corpus
- Entity tuples and roles within relationship (direction)
  - Who killed whom?
- Entity tuples and relationship type
  - Simplest: Verb of the sentence between entities
  - More advanced: Verb combining subject (E1) with object (E2)
  - But also nouns (interaction) and adjectives (interacting) can express relations
- Modifier of a relationship
  - Hedging: Might, could, should, not, ...

Is it Hard?

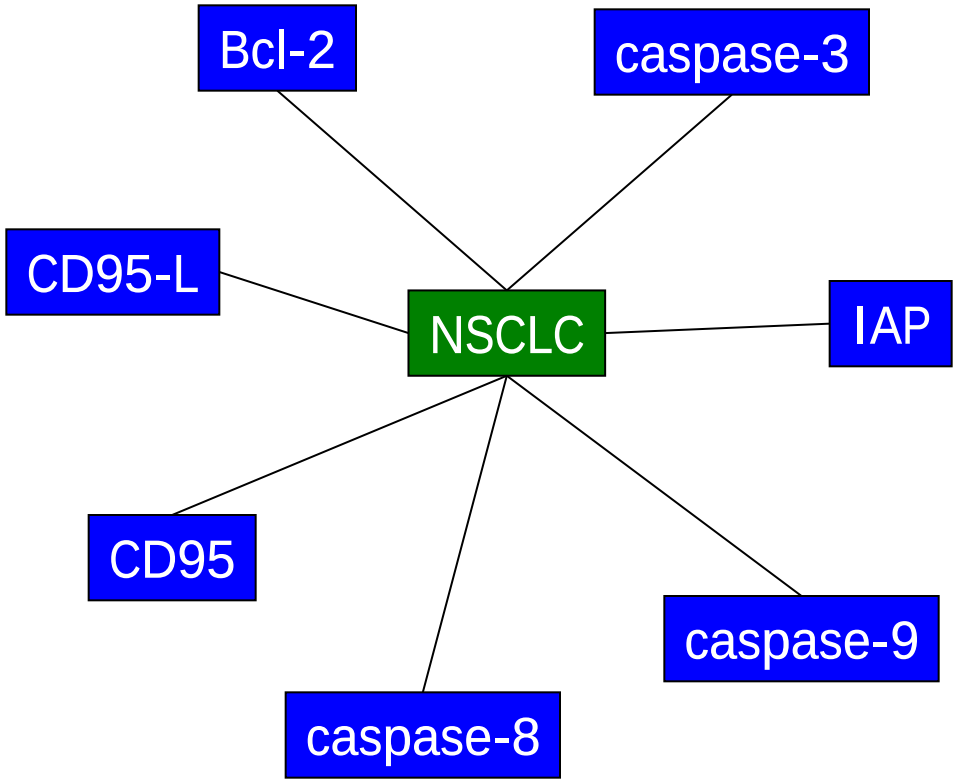
- Recognizing entities is difficult
  - Assume precision=0.8 for NER
  - Then, even a perfect binary RE has expected quality of only 64%
    - Currently large interest in joint inference (NER+RE in one step)
  - The higher the arity of the relationship, the worse
  - Often, RE is evaluated on a corpus pre-annotated with entities
- Sentences may contain more than one relationship
- Relationships may span sentences (coreference resolution)
- Enumerations in sentences (and, or)
  - “Oracle bought MySQL and RDB, while MySQL previously bought Adabas, which was then re-bought by SAP”
  - “TF-a must up-regulate RAS or b-RAF to induce this behavior“

Content of this Lecture

- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
  - Protein-Protein-Interactions

RE using Co-occurrence

„NSCLC often becomes resistant to chemotherapy due to multiple defects found in expression of CD95-L, CD95 and members of the Bcl-2 and IAP family, as well as caspase-8, -9 and -3 as examined by immunohistochemistry, ..“



Co-occurrence: 28 relationships, 21 false positives

Co-Occurrence-based RE (co-RE)

- Appearing together in a context
  - A sentence, a paragraph, a window of n words
  - Larger context: Higher recall (even across sentences), lower precision
  - Best context size for a given relationship can be learned
- General, co-RE yields high recall yet poor precision
  - Problems with enumerations, nested structures, long sentences, ...
  - Completely agnostic to relationship type
- Improvement: Pre-filtering sentences for “type’ness”
  - For instance, filter by a set of verbs or trigger words
- A fine-tuned co-RE often is quite a challenging baseline

Content of this Lecture

- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
  - Protein-Protein-Interactions

Pattern-Based Approaches to PPI Extraction

- Language pattern
  - Sentence
    - ... GENE regulates expression of GENE ...
    - ... GENE is strongly suppressed by GENE ...
  - Adding part-of-speech
    - ... GENE VRB NOM PRP GENE ...
    - ... GENE is ADJ VRB PRP GENE ...
- Different levels of generality
  - ... GENE .* VRB .* GENE
    - Simple rules, high recall, low precision
  - ... GENE [is] ADJ? {regulat|suppres} NOM? PRP GENE
    - Complex rules, lower recall, higher precision
- Balanced precision/recall requires many rules

State-of-the-Art

- Most systems work on hand-crafted sets of pattern
  - Hundreds of pattern
  - Enormous effort
  - Need to be created for any type of relationship
    - Protein-protein, gene-disease, disease-drug, ...
- One idea: Learn patterns from weakly labeled data
  - Semi-supervised learning

AliBaba Workflow (Hakenberg et al. 06, 07, 08, 09)

PubMed

IntAct

Protein pairs

Search sentences

Linguistic annotation

Initial patterns

Clustering

Consensus pattern

Alignment

Extracted PPI

Initial Pattern – Distant Supervision

- Extract all pairs of proteins from IntAct
  - Only the names, not the evidence / links
  - Gold standard: These interactions are assumed to be real
- Find all sentences in PubMed
  - Pair of IntAct-proteins and “interaction word”
  - “... FADD immediately activates procaspase-8 ...”
- Extract core phrases
  - Width: Parameter
  - “...show that FADD *immediately activates* procaspase-8 during...”
- Annotate with linguistic information

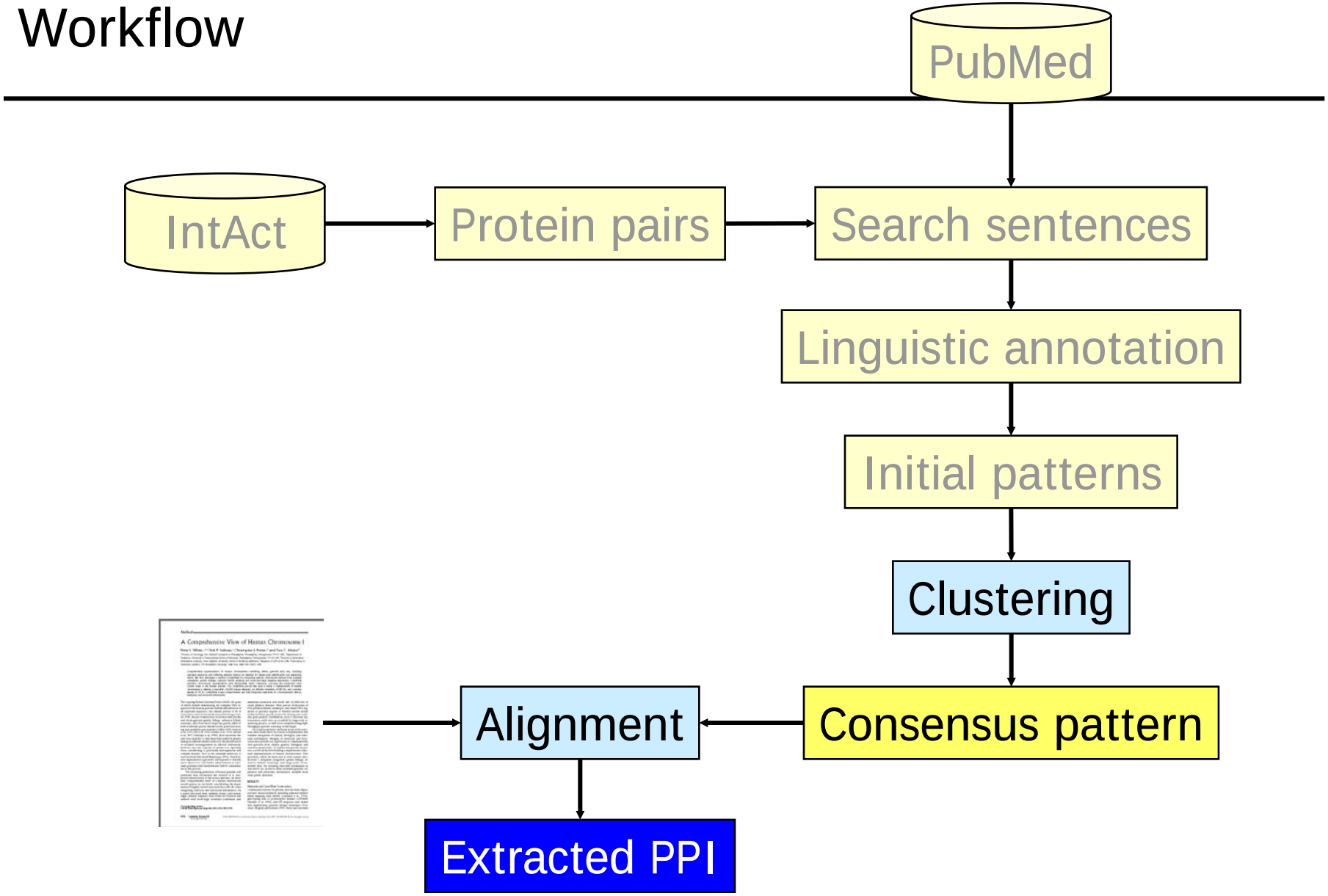
Linguistic Annotation

- Multi-layered pattern

Original token	FADD	immediately	activates	procaspase-8
Class / POS	PTN	ADV	VRB	PTN
Word stem	PTN	immediat	activat	PTN

- Initial pattern set
  - Highly specific
  - Can be used immediately, but results in very low recall
- Generalization
  - Find clusters of similar patterns
  - For each cluster, generate consensus pattern

Workflow

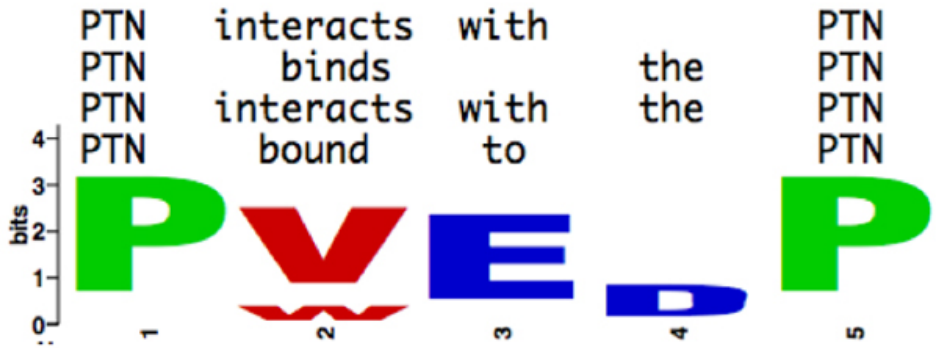

PubMed
IntAct
Protein pairs
Search sentences
Linguistic annotation
Initial patterns
Clustering
Consensus pattern
Alignment
Extracted PPI

Clustering and Generalization

- Distance matrix for all pairs of initial patterns
- Hierarchical clustering
- Build consensus pattern using multiple sentence alignment

P_1	PTN	SYM	PTN	IVBD	PTN
P_2	PTN	CC	PTN	IVBD	PTN
P_3	PTN	SYM	PTN	IVB	PTN
P_4	PTN	CC	PTN	IVBD	PTN
P_5	PTN	CC	PTN	IVBD	PTN
P_c	$PTN_{5/5}$	$CC_{3/5}\|SYM_{2/5}$	$PTN_{5/5}$	$IVB_{1/5}\|IVBD_{4/5}$	$PTN_{5/5}$

Similarity of Language Patterns



- Sentence alignment

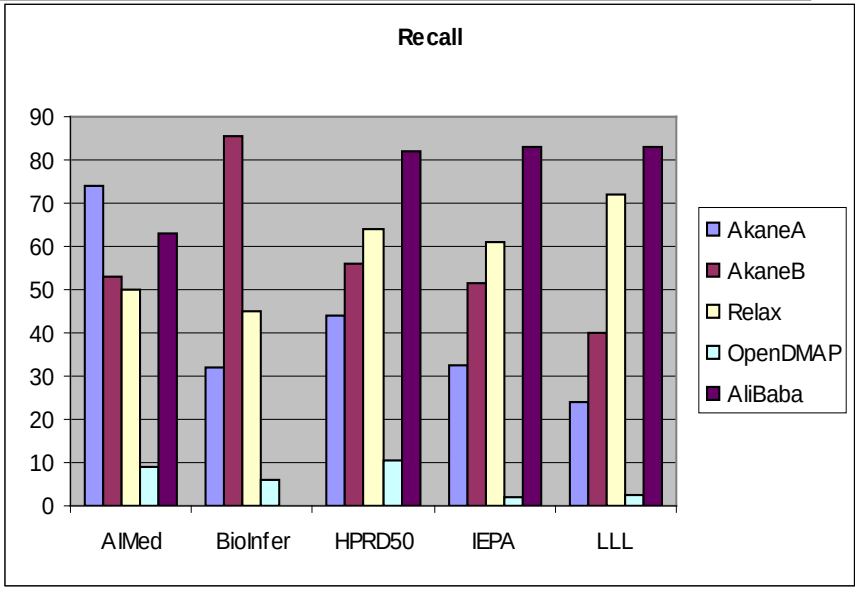
		NN	VBZ	DT	PTN	CC	PTN	IVBD	DT	PTN
	0	0	0	0	0	0	0	0	0	0
PTN	0	0	0	0	4	0	4	0	0	4
CC	0	0	0	0	0	5.6	0	0	0	0
PTN	0	0	0	0	4	0	9.6	0	0	0
IVBD	0	0	0	0	0	0	0	12.4	10.4	0
PTN	0	0	0	0	4	0	4	1.4	1.4	14.4

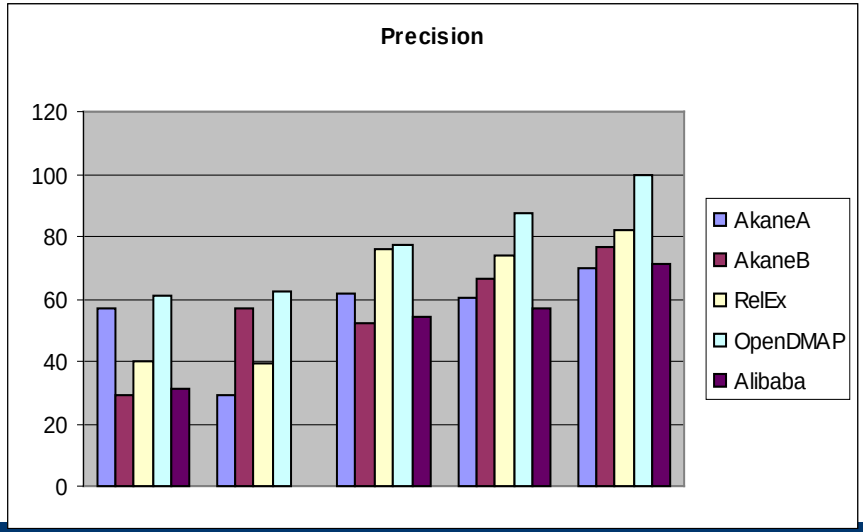
- Three-layer end-free alignment (token, stem, POS)
- Solved by dynamic programming

Tabelle mit Zahlen dazu, wie viele Pattern etc; es gab doch auch mal was für verchiedene Widths, ~~Kostenmatrizen etc.~~

Comparison (partly from Kabiljo et al. 09)

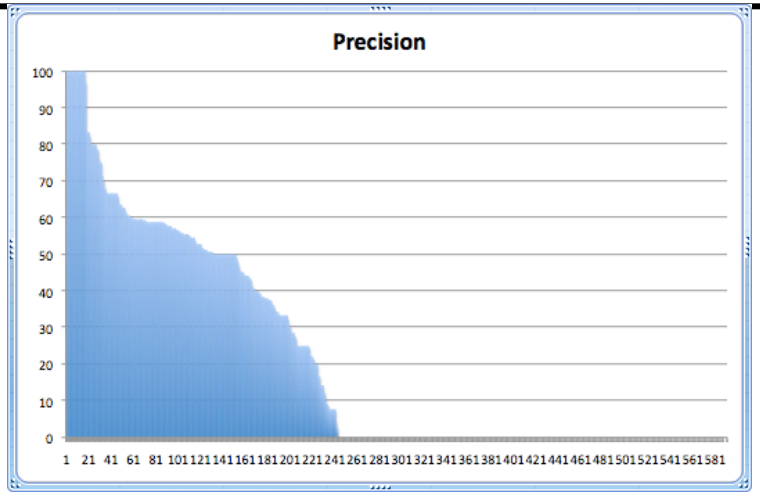
- Some results
  - AliBaba: Very good recall, acceptable precision
  - OpenDMAP: Very good precision, very low recall
  - RelEx: Best in F-measure
- Our advantage
  - Patterns are learned automatically
  - Simple tuning towards higher precision / higher recall
  - Adaptable to new problems



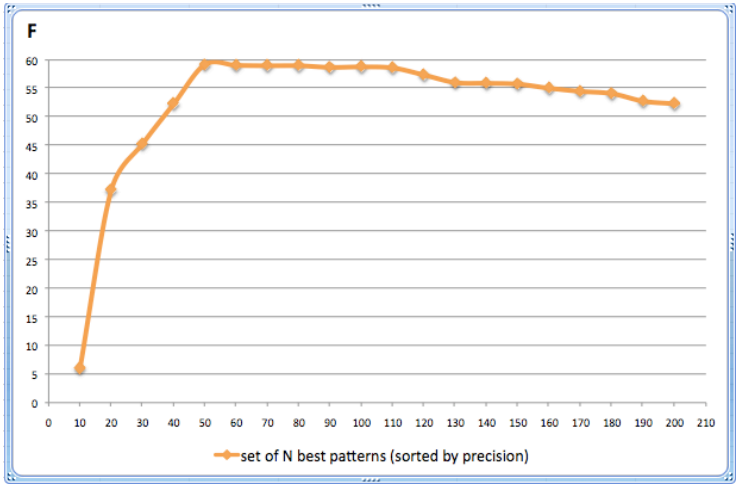


Good and Bad Patterns (BioNLP09)

- Large differences in the quality of individual patterns



- Using only the best pattern



Bootstrapping – Alternative to weak supervision

- Systems like AliBaba require a set of positive pairs as input
- These might not always be available in large quantities
  - Or in satisfying quality
- Bootstrapping
  - Start with a small set of high quality pairs
  - Apply to corpus and rank all extracted relations by confidence
  - Add relations with highest confidence to the set of positive pairs
  - Systems: Dare [XUL08], SnowBall [AH00], TextRunner [BCS+07]
- The trick is the scoring of extracted data
  - Use confidence of the extraction algorithm, number of times a particular pair is extracted, background knowledge, ...
  - Choosing the wrong relationships creates more and more garbage
    - Semantic drift increases after each iteration

Content of this Lecture

- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
  - Protein-Protein-Interactions

Classification-based Relationship Extraction

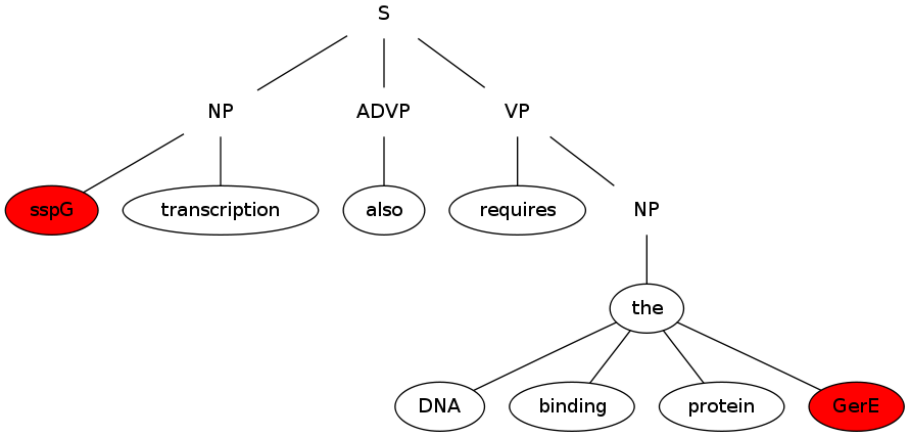
- Idea: Classify each pair of entities
  - Consider each entity pair (in a sentence) as an object
  - Compute a feature vector for this object
    - POS tags, distance, words, words in between, path in the dependency tree connecting the two, neighborhood, trigger words, ...
  - Learn a model from training data
  - Classify each object as having the relationship or not
- Any classification method can be used
- Finding the right features is essential

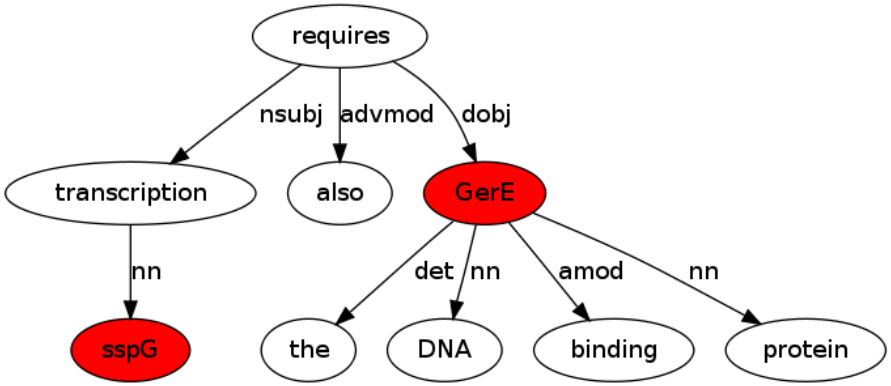
- As always in ML: Beware of overfitting

Representations of a Sentence

SsgG transcription also requires the DNA binding protein GerE

sspG	PROTEIN
transcription	NN
also	RB
requires	VBZ
the	DT
DNA	NN
binding	NN
protein	NN
GerE	PROTEIN



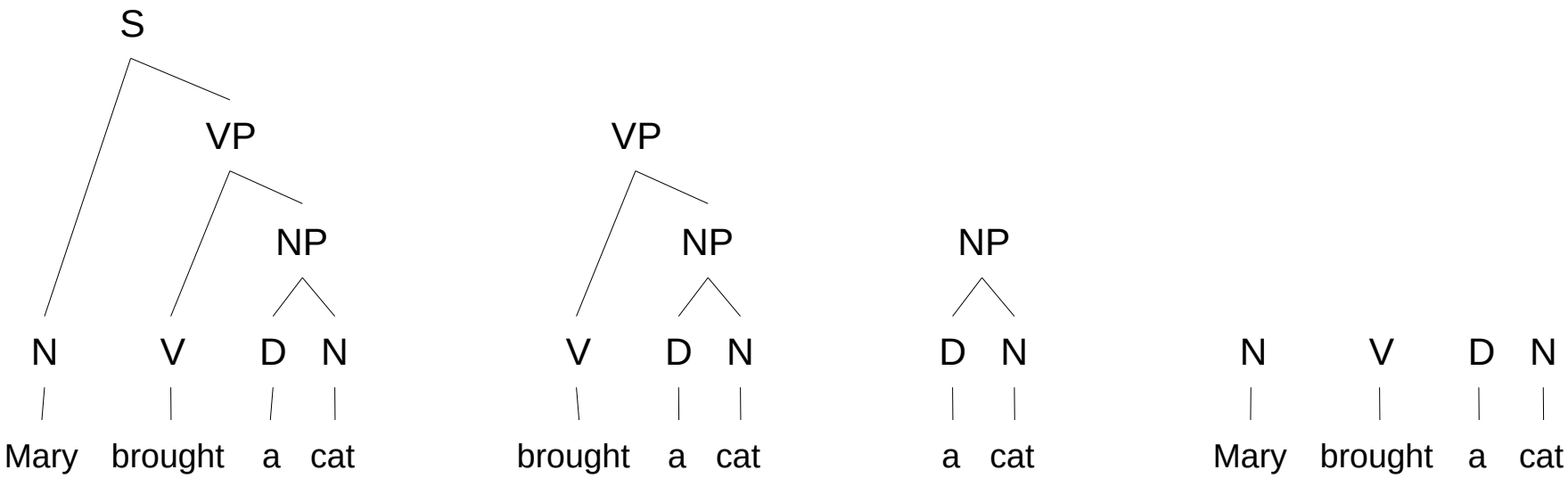


SVMs and the Kernel Trick

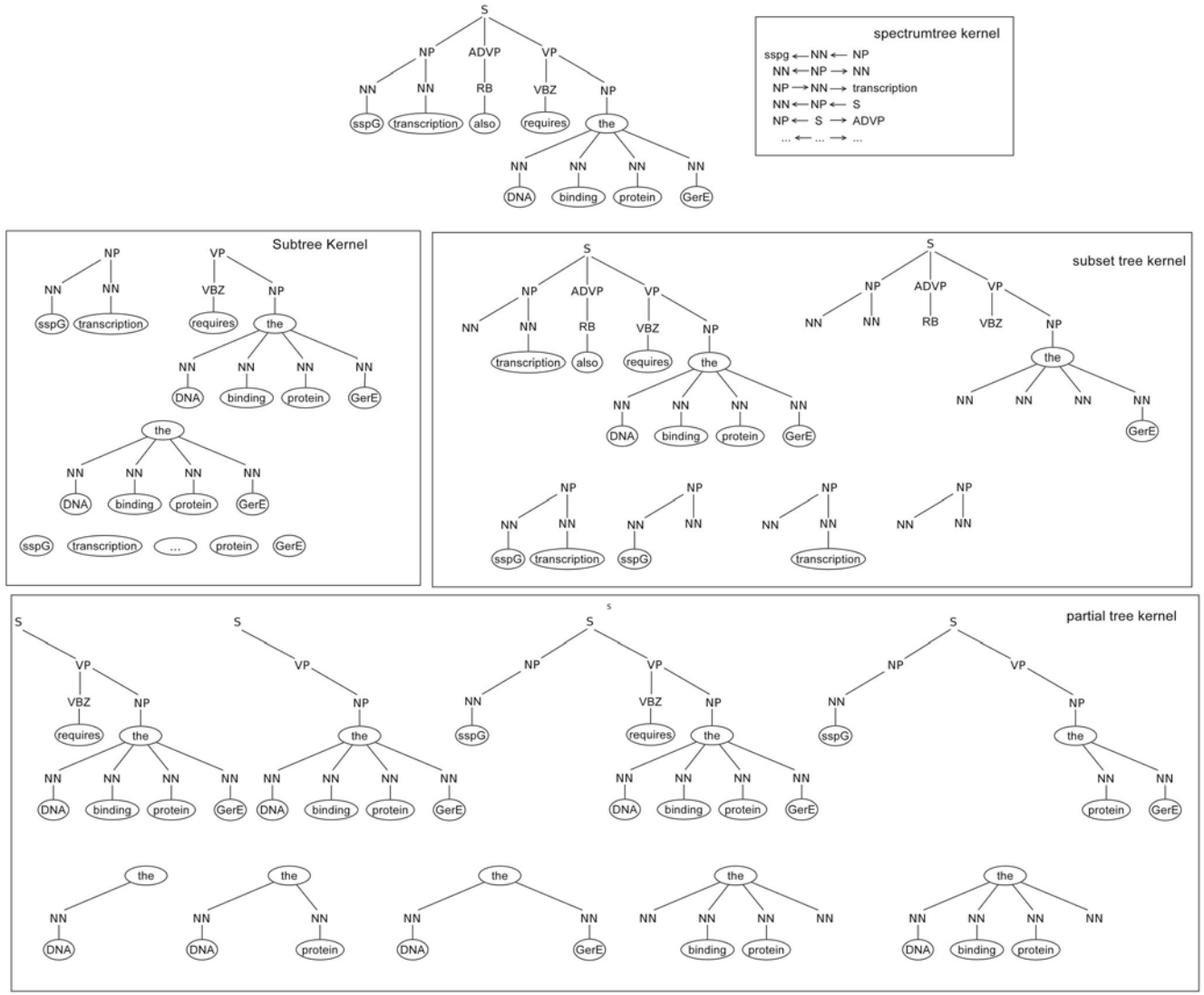
- How can we represent dependency trees in a feature vector such that similar trees lead to similar vectors?
- Elegant way: Kernel Trick
  - The learning problem in SVMs can be rewritten such that objects need not be explicitly described by features
  - Instead, one has to define a Kernel function computing the similarity of two objects
  - This function (and the object representations) is treated as a black box by the SVM
- We need a similarity measure for trees

Convolution Kernels

- General idea: Measure similarity of dependency trees in terms of common substructures
- One idea: All subtrees
  - Compute all subtrees of both objects, then use SET-similarity
- Alternatives: All subgraphs, all edges, all ...



Convolution Kernels - representations



Tikk et al. 2010

Content of this Lecture

- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
    - Work by L. Döhling, partly based on S. Pietschmann
  - Protein-Protein-Interactions
    - Work by P. Thomas, D. Tikk, and I. Solt

Extracting n-Ary Relationships

- Option 1: Use co-occurrence
  - Whenever a sentence contains one entity of each requested type, extract the relationship
  - If for one type there are >1 entity: Chose closest (to what?)
    - Neglects grammar/semantic of sentences
  - If entities have a strong semantic relationship and are not highly ambiguous, this works quite well
- Option 2: Use n-ary patterns
- Option 3: Use classification
- Option 4: Map into many binary RE-problems
  - Compute binary RE's for each pair of the n-ary relationship
  - Generate n-ary relations (e.g. strategy of BioNLP'09 winning team)

Text Mining for the GFZ Earthquake Task Force

- Measures in case of an earthquake depend on the expected extend of damage
  - Here: Expected number of people injured / killed
- Early information typically is reported in news, but highly inconsistent and quickly changing
- Project: Find such information automatically
- Cast into a 5-ary RE problem
  - Who? (People, Students, ...)
  - How many? (many, some, 12, ten, ..)
  - What? (killed, trapped, injured, ...)
  - Negated? (not, ...)
  - Modifier for “how many”? (at least, more than, ...)

Example

- *“The death toll in an earthquake in south west China is now at least 32, with 467 injuries, media say.”*
  - [Who, How many, What, Negated, Injured]
  - [-, 32, death, -, “at least”]
  - [-, 467, injuries, -, -]

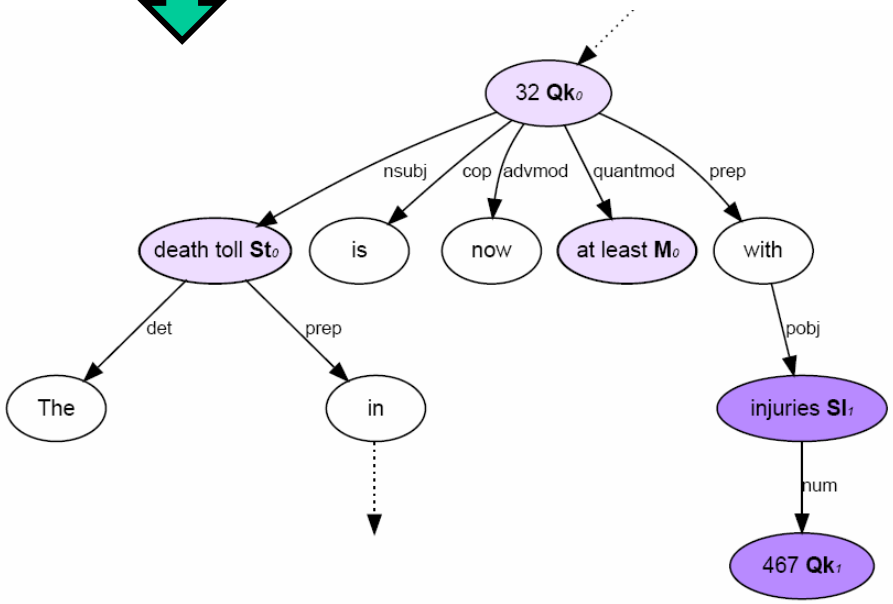
Approach

- Use word lists for Who? What? Negated? Modified?
- Use regular expression for “How many”?
  - Problem: Highly ambiguous, finds any number (problem for irrelevant texts)
- Learn paths in dependency trees between all pairs of entities from an annotated gold standard corpus
- Application
  - Identify all entities
  - Parse sentence
  - Extract paths
  - Match with learned paths
  - Extract binary relationships

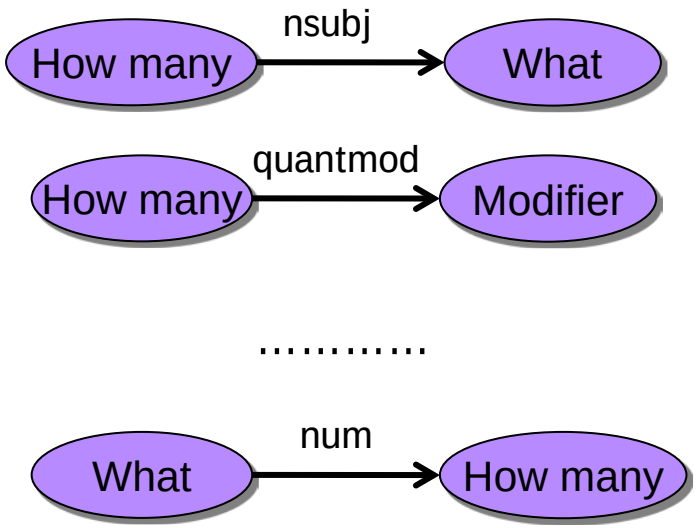
Binary to 5-ary Rels.

The death toll in an earthquake in south west China is now at least 32, with 467 injuries, media say."

Dependency graph



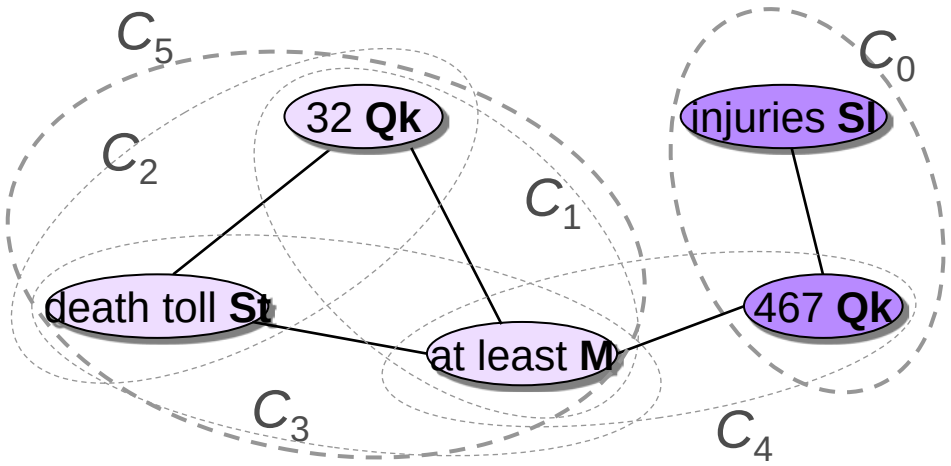
396 pattern:



From Binary to 5-ary Relationships

- Build graph representation from extracted binary relations
  - Find maximal cliques

The death toll in an earthquake in south west China is now at least 32, with 467 injuries, media say."



Many Further Tricks

	BestConfigP	BestConfigR	BestConfigF1
IgnoreCase4NER	–	+	–
UseStem4NER	–	+	–
Dependenzschema	Collapsed	Basis, CCprocessed	Basis
IgnoreCase4RE	*	–	*
UseStem4RE	+	–	*
UsePOS4RE	–	+	–
IgnoreEntitySubtype	+	+	–
IgnoreDepDirection	–	+	+
IgnoreDepType	–	+	+

	RE			
	P	R	F1	FP/TP/FN
Standard	.752 [.667;.823]	.495 [.423;.568]	.597 [.527;.664]	31/94/96
BestConfigP	.793 [.715;.855]	.563 [.484;.638]	.658 [.589;.722]	28/107/83
BestConfigR	.523 [.459;.586]	.711 [.629;.781]	.603 [.541;.660]	123/135/55
BestConfigF1	.765 [.690;.827]	.600 [.521;.672]	.673 [.607;.732]	35/114/76

Content of this Lecture

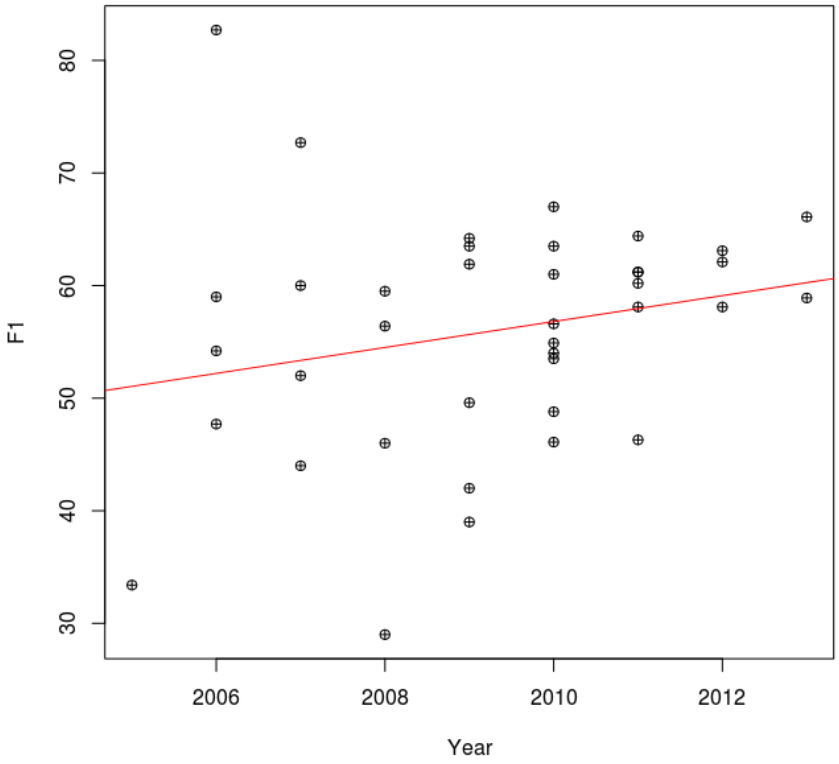
- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
    - Work by L. Döhling, partly based on S. Pietschmann
  - Protein-Protein-Interactions
    - Work by P. Thomas, D. Tikk, and I. Solt

Convolution Kernels for PPI: Many Proposals

- Collins, M. and Duffy, N. (2001). Convolution kernels for natural language.
- Vishwanathan, S., Smola, A. (2002): Fast kernels on strings and trees
- Moschitti, A. (2006): Efficient convolution kernels for dependency and constituent syntactic trees.
- Kuboyama, T. et al. (2007). A spectrum tree kernel.
- Erkan, G. et al. (2007). Semi-supervised classification for extracting protein interaction sentences using dependency parsing
- Giuliano, C et al. (2007). Kernel Methods for Semantic Relation Extraction
- Airola, A. et al. (2008). All-paths graph kernel for protein-protein interaction extraction
- Palaga, P (2009). Extracting Relations from Biomedical Texts Using Syntactic Information, Magisterarbeit, HU Berlin
- ...

Cross-Validation – Published results

- More than 60 publications for PPI extraction over last years



Differences in Evaluation

- Single method has different results on different corpora
  - 19% on average (Annotation guidelines and pos/neg ratio)
- Gold-standard corpora are differently interpreted
  - 951 to 1071 positive and 4026 to 5631 negative instances
  - Self-Interactions are sometimes ignored
- Directed / Undirected relations
- Entity blinding is important requisite for new interactions
  - 3% points increase without entity blinding (Drug-Interactions)
- Cross-Validation type?
  - Pairwise cross-validation leads to 18% points overestimation in F1

Based on Pyysalo et al. „Why Biomedical Relation Extraction Results are Incomparable and What to do about it“

Differences – continued

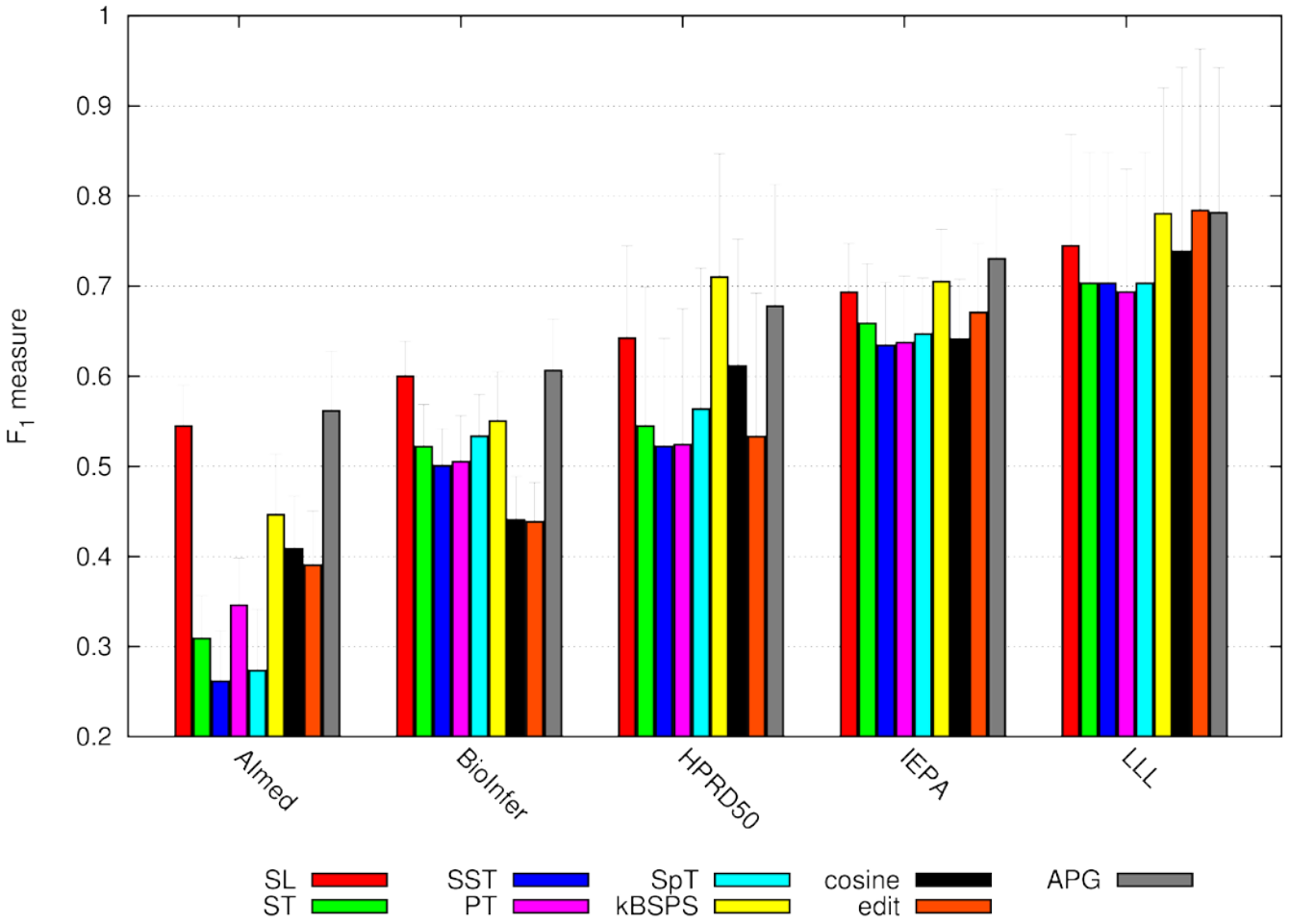
- How to build averages in cross-validation
  - Microaveraging (accumulate TP,FN,FP) or Macroaveraging (average precision/recall over ten folds)
- Exhaustive cross-validation with high dimensional parameter space
  - Identifies performance „spikes“
  - Large effect especially on smaller corpora
  - Estimate optimal threshold on test set
  - Ideal: Use test-corpus only once (e.g. BioNLP09-ST)

Based on Pyysalo et al. „Why Biomedical Relation Extraction Results are Incomparable and What to do about it“

Which One is the Best?

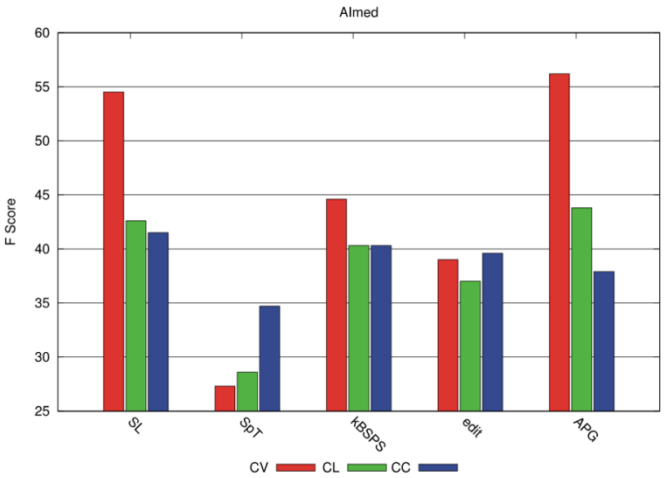
- Very difficult question
  - Different corpora, different evaluation schemes, different parsers, w/o protein identification, w/o parameter tuning, ...
- Reported results sometimes up to 90% F-measure
- Large-scale benchmark
  - 9 methods
  - 5 corpora
  - 3 evaluation schemes
  - Equal parser, equal treatment of NER, equal parameter tuning
- Bad news: "Real" performance remains unknown

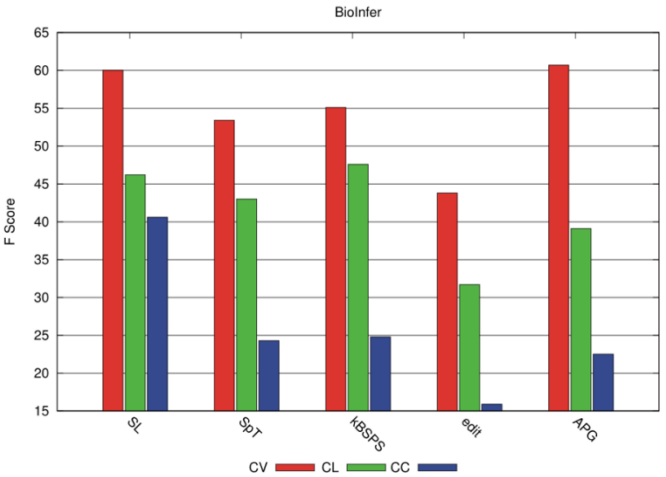
Cross-Validation (usual method)



But: Cross-Learning: ~10% drop in F1

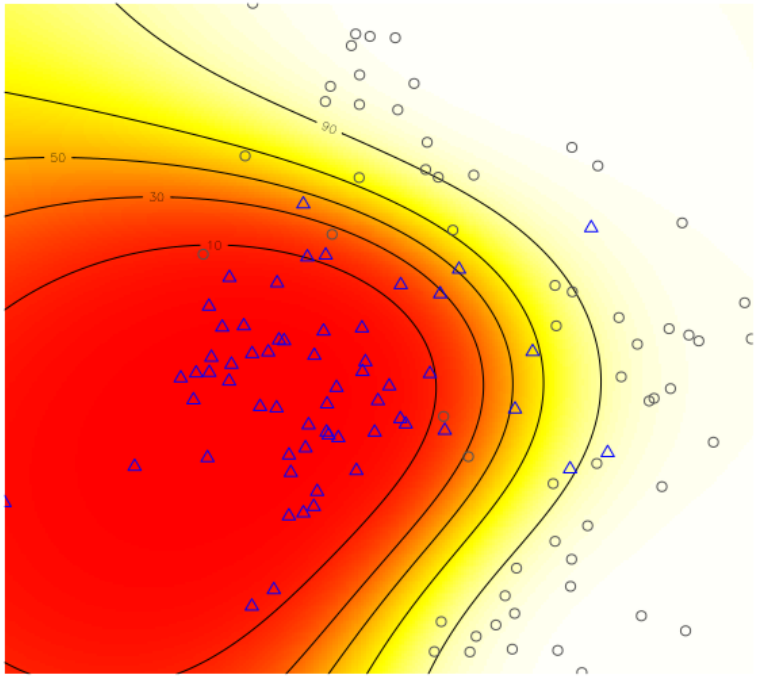
- CC probably overly hard
- CL: Best approximation of the real-case
- Some observations
  - APG generally best in CV setting, but not in CL / CC
  - SL on par with best methods, though using only POS tags
  - kBSP quite good on BioInfer, but not on AIMed
- In CL/CC, simple pattern-based methods perform equally well

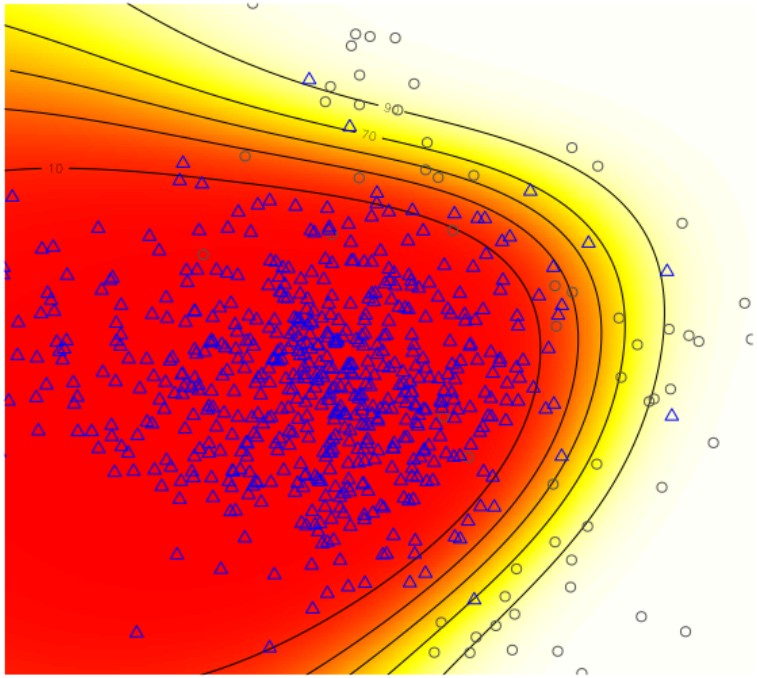




Classifier tend to predict majority class

- Sample from the same distribution
  - Balanced/Unbalanced data set and learn a classifier





Classifier tend to predict majority class

- Remove presumably negative instances
  - Rule1: Two entities use the same mention
  - Rule2: Both entities have anti-possesive governors w.r.t. the relation (generated on training set)
  - Rule3: Entity2 is an abbreviation of Entity1
- Leads to:
  - Better balanced pos/neg ratio
  - Faster runtimes
  - Improved F1 for all five corpora

Chowdhury et al. 2012, „Impact of Less Skewed Distributions on Efficency and Effectiveness of Biomedical Relation Extraction“

Conclusions

- Unbiased evaluation of ML-based method reveals 5-20% performance drop compared to CV setting
- Highly-tuned ML-based methods not (much) better than “simple” pattern matching
- Large differences between corpora: Extrapolation of performance to new text is very questionable
- Dependency-tree based methods not (much) better than best ones using POS information
- Still: Three methods are best (APG, JSRE/SL, KBSP)
  - And JSRE is by-far the fastest
- A large corpus for less biased evaluations is still missing
- Field should focus on more specific questions