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 – not here

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 to each other
  - Output: Tuples (mostly pairs) of entities with predefined semantics
  - Semantics often only implicitly defined through training corpus
- Entities and roles within relationship (direction)
  - Who killed whom? Who bought whom?
- Entities and relationship type
  - Detect entities and deduce semantics of their relation
  - Simplest: Verb of the sentence containing the entities
  - More advanced: Verb combining subject (E1) with object (E2)
  - But also nouns (interaction) and adjectives (interacting) can express semantics
- Modifier of a relationship
  - Hedging: Might, could, should, not, …

Is it Hard?

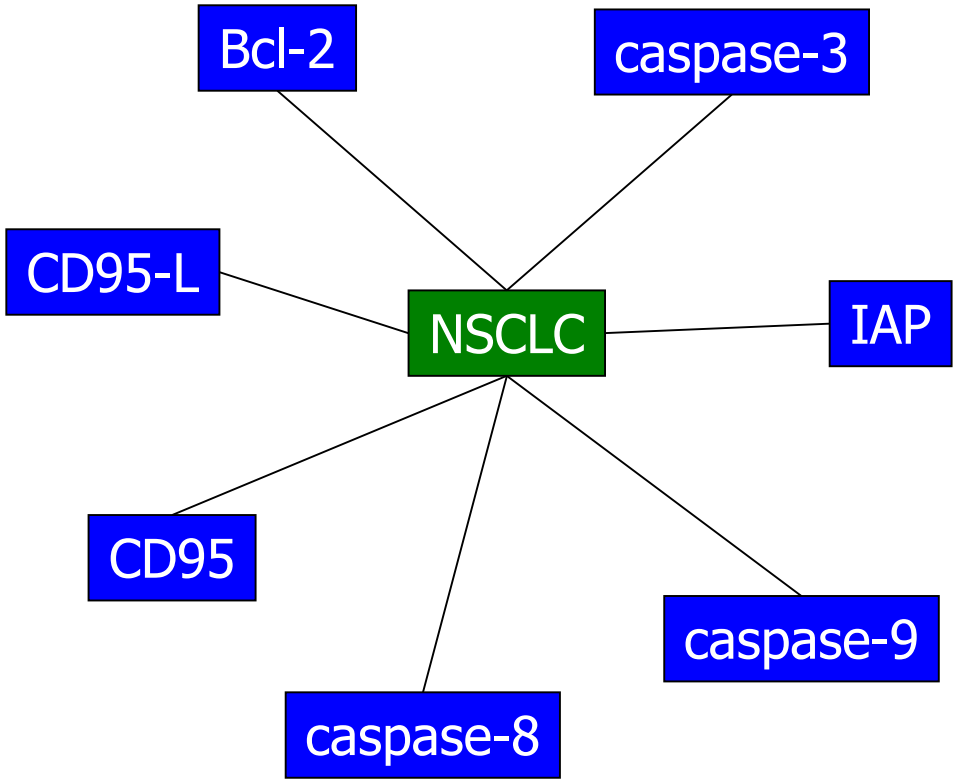
- Recognizing entities is difficult
  - Assume a recall of 80% for NER
  - Then, even a perfect binary RE has expected recall of only 64%
  - The higher the arity of the relationship, the worse
    - … the recall; but precision often increases
  - Often, RE is evaluated on a corpus pre-annotated with entities
- Sentences may contain more than one pair / relationship
- Relationships may span sentences (co-references)
- 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)

- All pairs of entities appearing together in a context
  - A sentence, a paragraph, a window of n words
    - Larger context: Higher recall (e.g. across sentences), lower precision
  - Best context size for a given relationship can be learned
    - Typical distance of entities having this relationship
- 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 is a reasonable baseline
  - Filter documents (topical classification); filter sentences (verbs); special treatment of conjunctions (replace with a single entity)

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 RE (see rule-based NER)

- Language pattern (aka "Hearst Pattern")
  - Look at words occurring in sentences expressing a relationship
    - … GENE regulates expression of GENE …
    - … GENE is strongly suppressed by GENE …
  - Adding part-of-speech
    - … GENE VRB NOM PRP GENE …
    - … GENE is ADV VRB PRP GENE …
- Different levels of generality
  - … GENE .* VRB .* GENE
    - Simple rule, high recall, low precision
  - … GENE [is] ADV? {regulat|suppres} NOM? PRP GENE
    - Complex rules, lower recall, higher precision
- Balanced precision/recall requires many rules

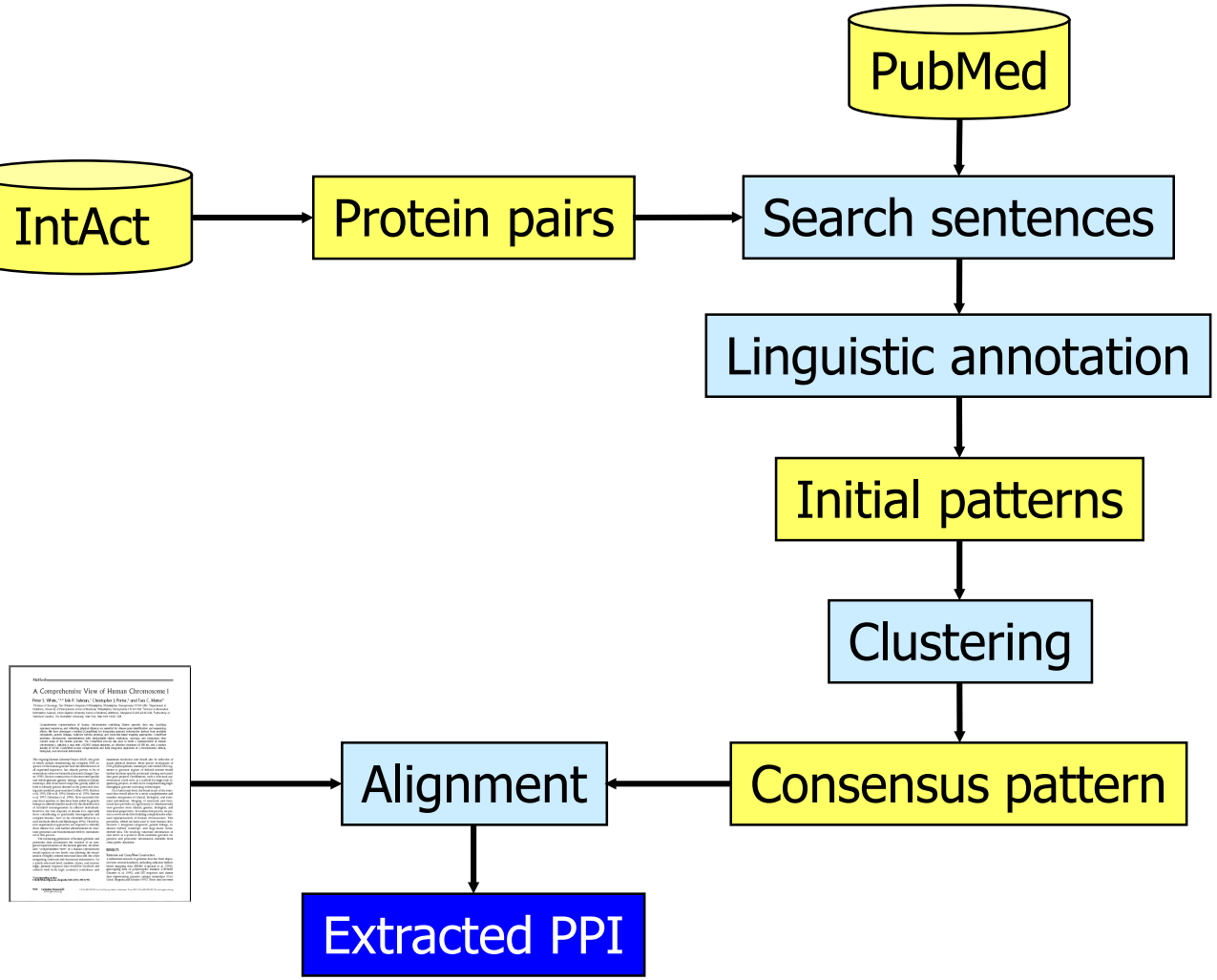
State-of-the-Art

- Most pattern-based systems work on hand-crafted sets of pattern
  - Recall: Users love pattern/rule-based approaches
  - Good recall quickly requires hundreds of pattern – large 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
  - More specific term: Distant supervision
    - Learn from training data that is different than what you intend
  - User-friendly: Patterns can be inspected, removed, modified, …

Idea

- Assume we seek protein-protein-interactions (PPI)
- Fortunately, there exist databases of PPIs, e.g. IntAct
- Hypothesis: If a pair of proteins known to interact (from IntAct) co-occur in a sentence, then this sentences expresses a PPI
  - That's a strong assumption! Certainly often wrong
- Can be used to quickly find thousands of PPI-carrying sentences
- Groups of sentences are then turned into patterns
- These patterns can be matched against new text to find novel PPI

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



Initial Pattern

- Extract all pairs of proteins from a PPI database
  - Only the names, not the evidence / links
  - All these interactions are assumed to be real
- Find all sentences in PubMed with a pair and an "interaction word"
  - "… FADD immediately activates procaspase-8 …"
- Annotate with linguistic information
- Extract core phrases
  - Width: Parameter
  - "…show [that FADD *immediately activates* procaspase-8 during] …"

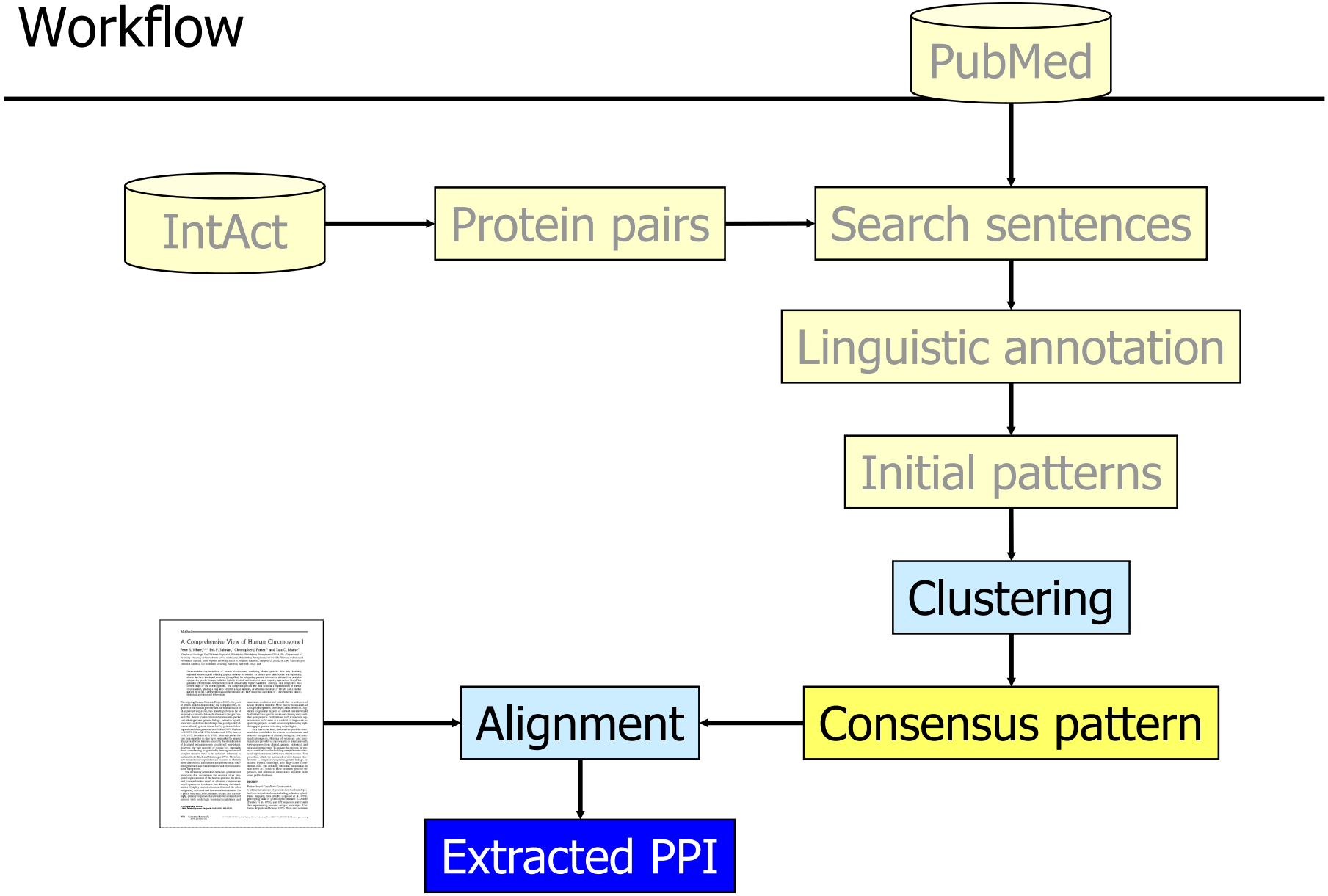
Linguistic Annotation

- Multi-layered pattern abstracting from concrete entities

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

- Initial pattern set, one from each matching sentence
  - Highly specific
  - Can be used immediately, but results in low recall
  - Finds mostly the sentences they were generated from
- We need to generalize
  - 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	$PTN_{5/5}$	$CC_{3/5}\|SYM_{2/5}$	$PTN_{5/5}$	$IVB_{1/5}\|IVBD_{4/5}$	$PTN_{5/5}$

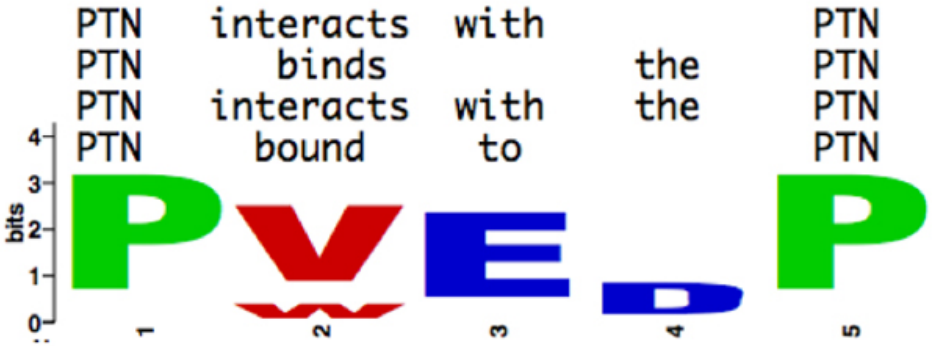
Similarity of Language Patterns

- What is the "distance" between two multi-layer pattern?
- Many notions of distance are possible (e.g. Jaccard)
- We use sentence alignment
  - Find the minimal set of operations (insert, delete, rename token) that transforms one sentence into the other
  - The size of this set is used as distance (=edit distance)
  - Can be solved efficiently using dynamic programming
  - Slightly more complicated due to the three layers of a pattern

		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

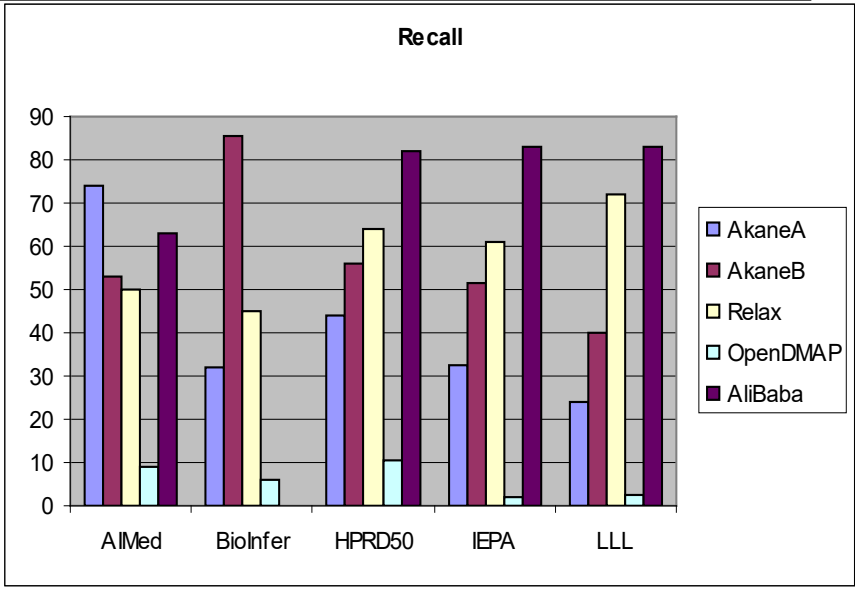
Building Consensus Pattern

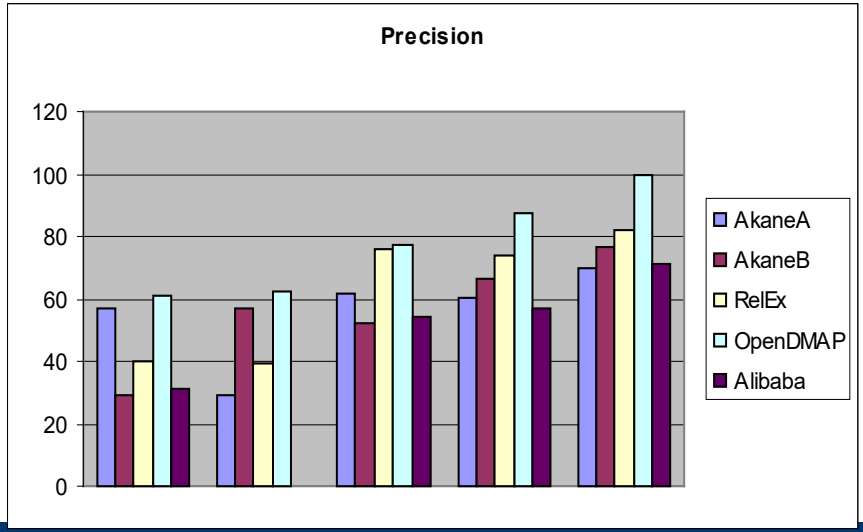
- A cluster consists of k patterns (of poss. different lengths)
- Many ways to find a consensus (e.g., the median pattern)
- We use multiple pattern alignment (MPA)
  - Arranges all patterns in a table such that the least number of empty cells and none-pure columns emerge
    - Dynamic programming
    - But: Exponential in k
  - Use greedy approximation
- Each MPA is turned into a pattern
  - The pattern is as long as the MPA
  - In each position, it defines weights for matches according to the distribution of values in the MPA



Comparison (partly from Kabiljo et al. 09)

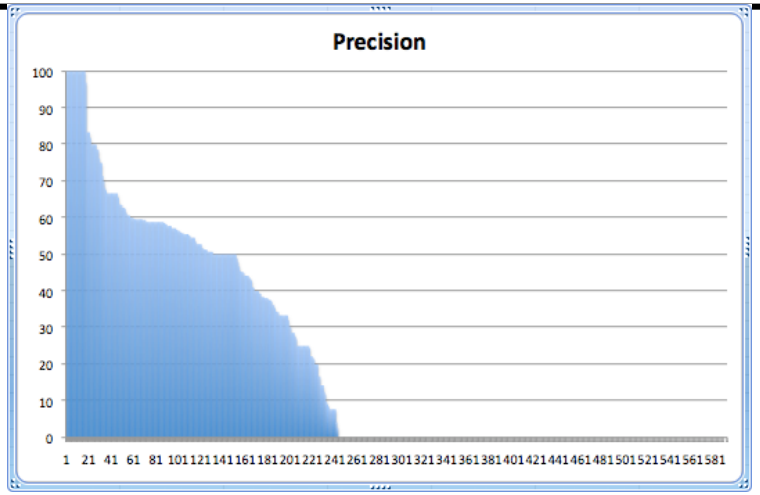
- Some results
  - AliBaba: Very good recall, acceptable precision
  - OpenDMAP: Very good precision, very low recall
  - RelEx: Best in F-measure
    - RelEx uses dependency trees
- Our advantage
  - Patterns are learned automatically
  - Simple tuning towards higher precision / higher recall
  - No annotated corpus necessary



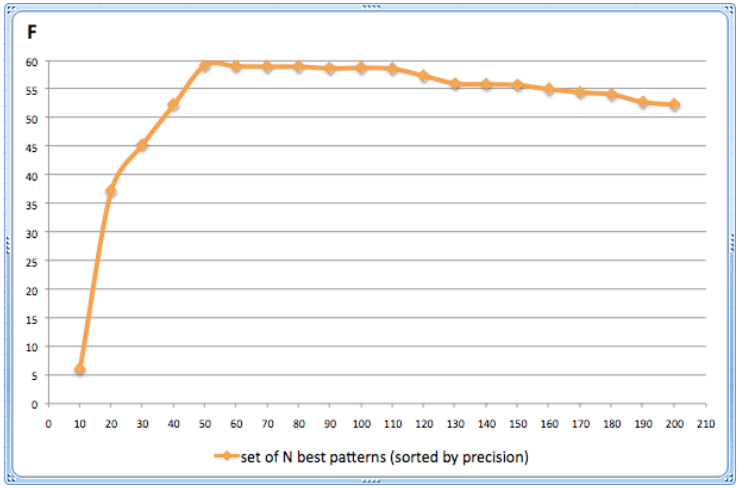


Good and Bad Patterns (BioNLP09)

- Large differences in the quality of individual patterns



- Using only the best pattern
  - Needs a training corpus



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 relations
  - 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

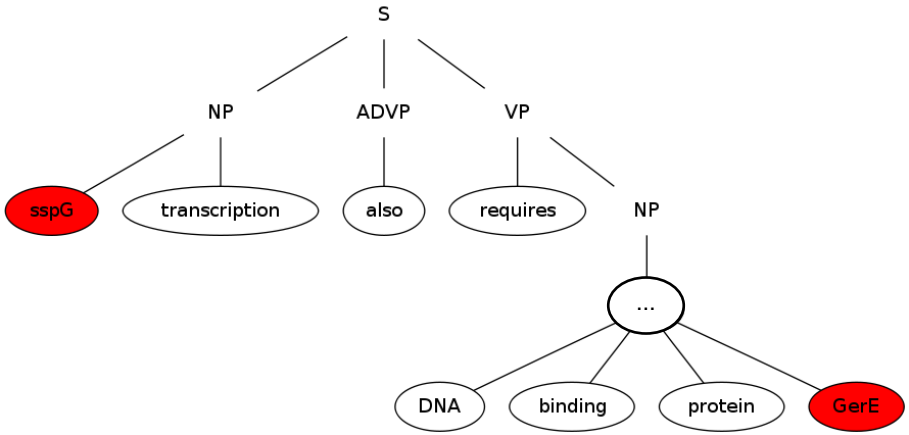
Classification-based Relationship Extraction

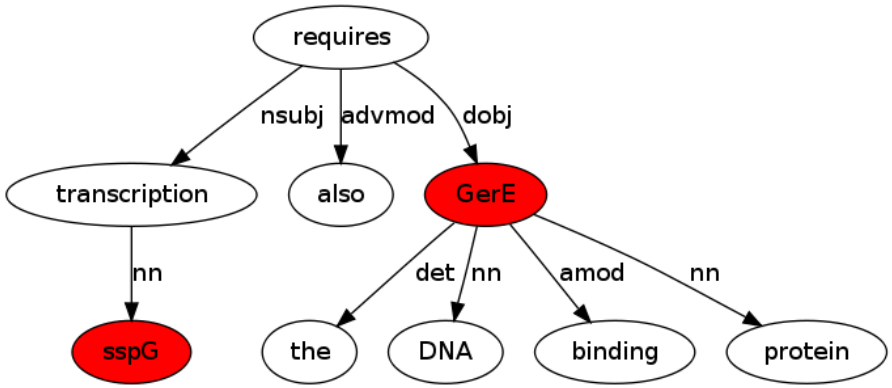
- Idea: Classify each pair of entities
  - Consider each entity pair (in a sentence) as an object
  - Compute a feature vector for this object
    - POS tags, distance, words, words in between or in the neighborhood, presence of trigger words, …
  - Learn a model from training data
  - Classify each object as having the relationship or not
- Any classification method can be used
- Finding the best features is essential
- As always in ML: Beware of overfitting
- But: How to integrate structured information?
  - Especially: Syntax trees, dependency trees

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





Structured Features

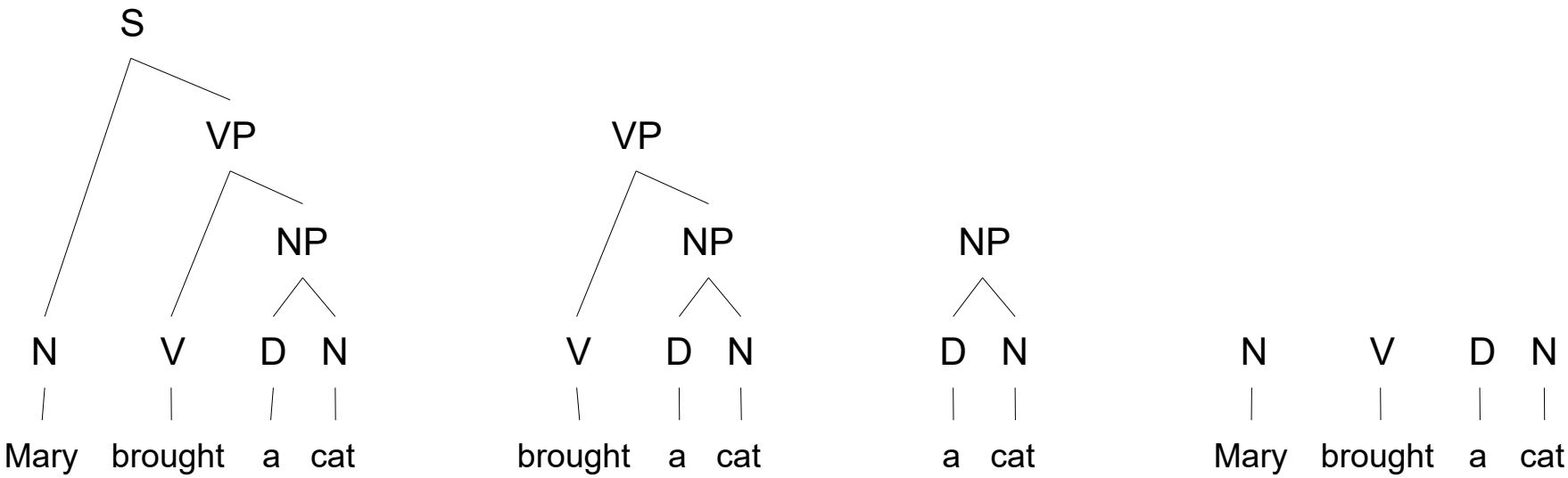
- How can we turn trees-like data to feature values?
- Very hard
  - There are enormously (exponentially) many possible paths / subtrees
  - This would lead to exponentially many features
  - This feature vector will be extremely sparse – most values will be seen only once
    - Which makes ML impossible
- Classical features do not encode similarity of values
  - Is the dimension "king" similar to dimension "queen"?
- We need to bring similarity functions into classification
  - Note: This is no problem for k-Nearest-Neighbor classifier

SVMs and the Kernel Trick

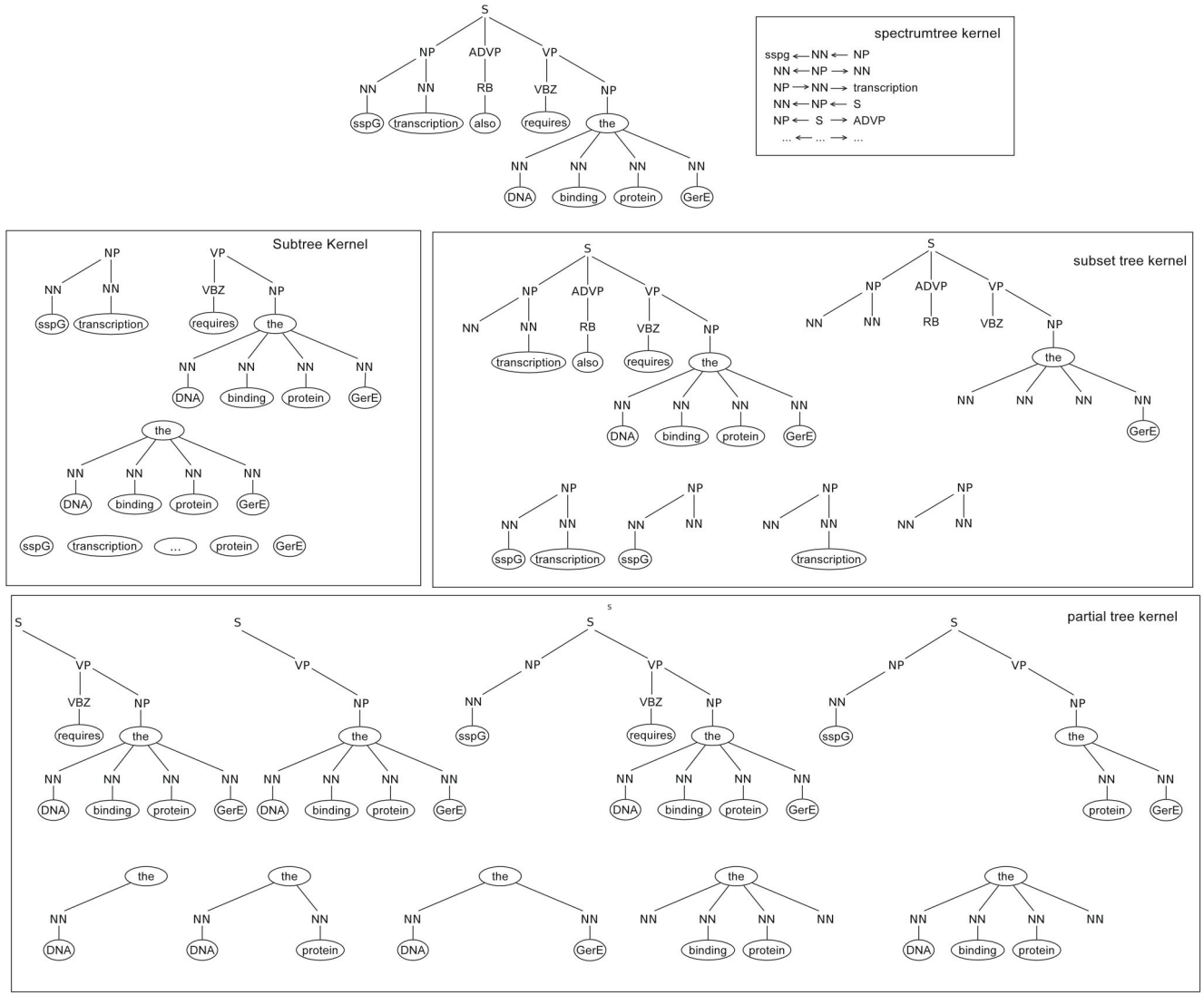
- Elegant way: Kernel Trick
  - The learning problem in SVMs can be rewritten such that objects need not be explicitly described by features
  - Instead, one has to define a Kernel function computing the similarity of two objects
    - This similarity does not need to be based on features
  - This function (and the object representations) is treated as a black box by the SVM
- Open: We need a similarity measure for trees

Convolution Kernels

- General idea: Measure similarity by counting 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

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 and aggregate 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, -, -]

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
    - Locations are easily assigned a role in a relationship, numbers not
- 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
  - Aggregate into n-ary relations

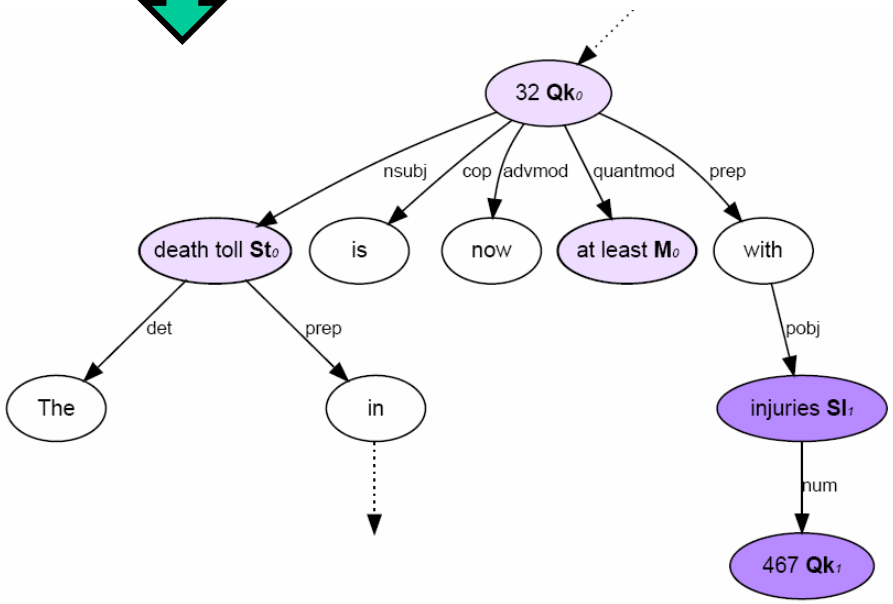
Equator Approach [Döhling, Leser, 2014]

- Entity identification
  - Word lists for Who? What? Negated? Modified?
  - Regular expression for "How many"?
    - Problem: Highly ambiguous, finds any number, many matches
- Binary relationships
  - Learn paths in dependency trees between all correct pairs of entities within a gold standard corpus
- Aggregation
  - Assemble a graph from all binary relationships
  - Cliques in this graph are n-ary relationships

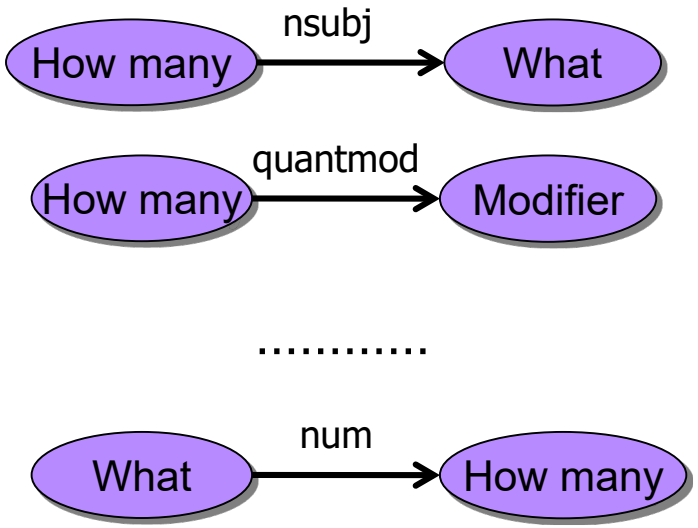
Binary to 5-ary Rels.

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

Dependency graph



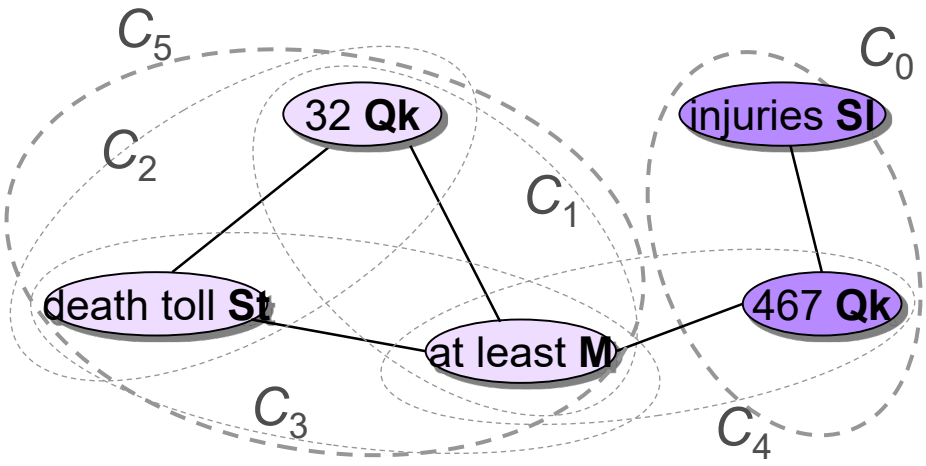
396 pattern:



From Binary to 5-ary Relationships

- Build graph 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

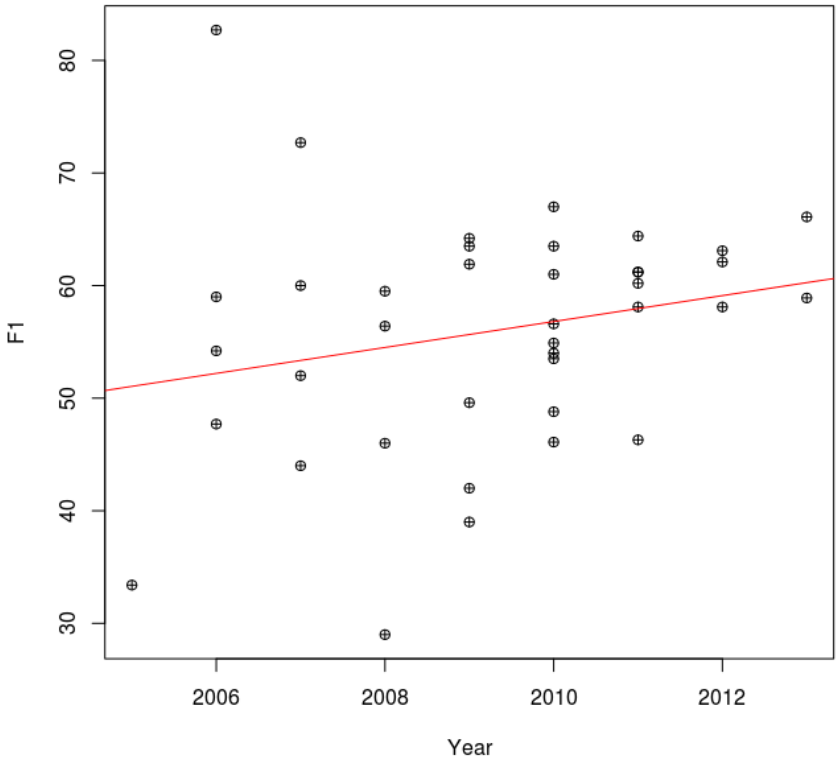
- Relationship Extraction
- Approaches
  - Co-Occurrence
  - Pattern-Based
  - Classification-Based
- Case Studies
  - Damage reports after an earthquake
    - Work by L. Döhling, partly based on S. Pietschmann
  - Protein-Protein-Interactions
    - Work by P. Thomas, D. Tikk, and I. Solt

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 [Thomas, 2015]

- More than 60 publications for PPI extraction over last years



Differences in Evaluation

- Single method has different results on different corpora
  - 19% diff on average
  - Many causes, such as diff annotation guidelines or 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 to find new interactions
  - 3% points increase without entity blinding (Drug-Interactions)
- Cross-validation type?
  - Which folds, how many?

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

Differences – continued

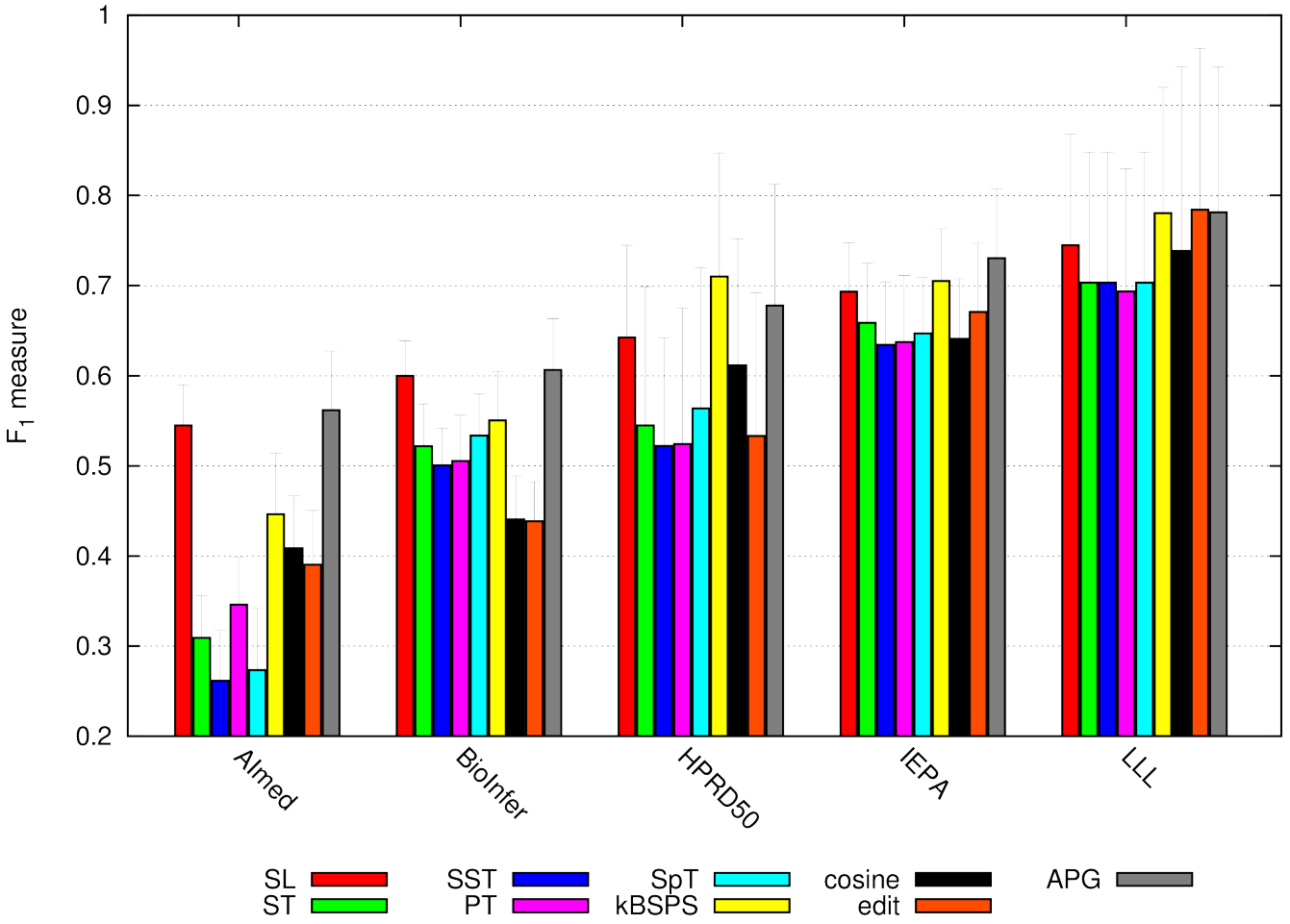
- How to build averages in cross-validation
  - Micro-averaging (accumulate TP,FN,FP of folds)
  - Macro-averaging (average P/R over folds)
- Obtaining hyper parameters: Parameter sweeps in high dimensional parameter space
  - Identifies performance „spikes"
  - Large effect especially on smaller corpora
  - Important (again): Use test-corpus only once

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

Which PPI-Extraction Method is the Best?

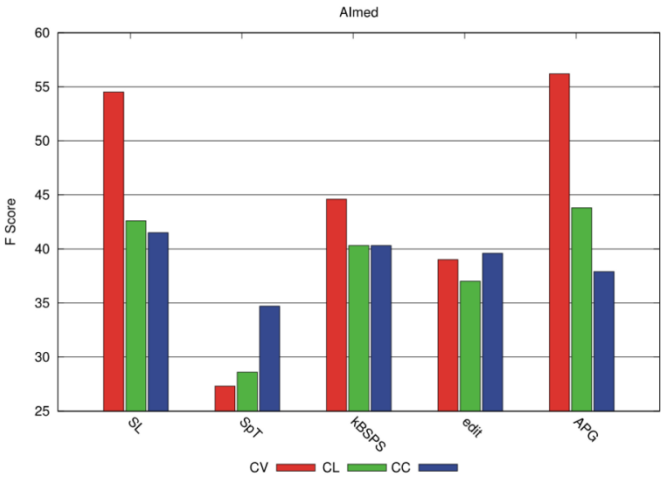
- Very difficult question
  - Different corpora, different evaluation schemes, different parsers, w/o protein identification, w/o parameter tuning, …
- Reported results sometimes up to 90% F-measure
- Large-scale benchmark
  - 9 methods
  - 5 corpora
  - 3 evaluation schemes
  - Same parser, same treatment of NER, same level of parameter tuning, same folds, same SVM, …

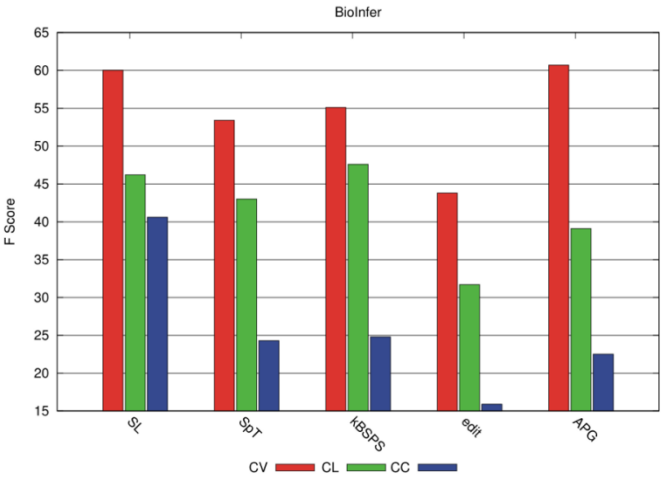
Within-Corpus Cross-Validation (usual method)



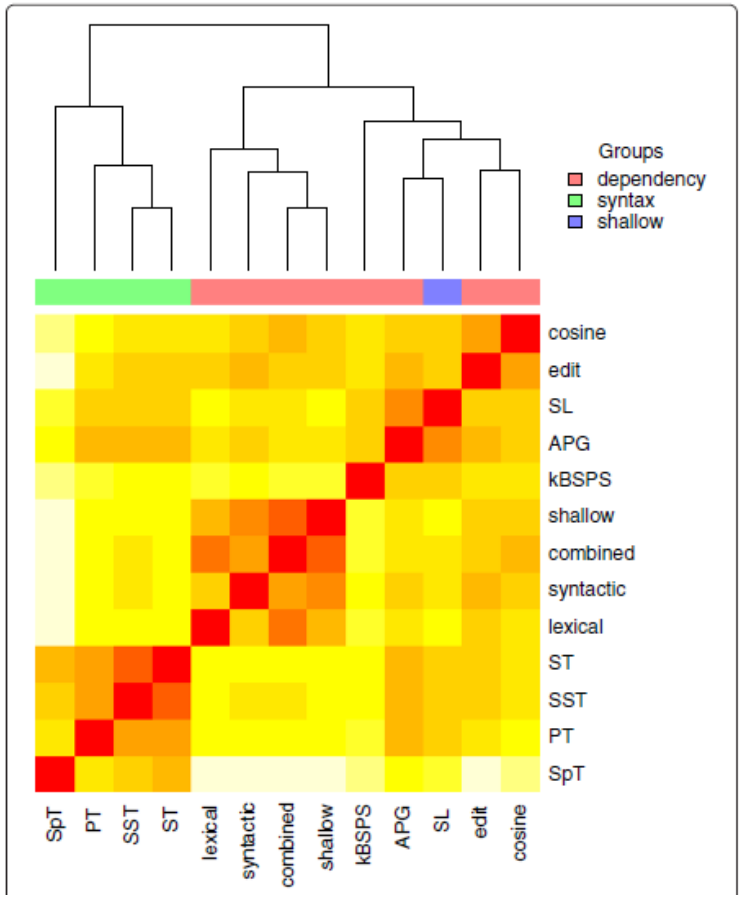
Cross-Learning: ~10% drop in F1

- Best approx. of the real-case
  - Learn on everything available except test data
- Observations
  - APG generally best in CV setting, but not in CL / CC (and very slow!)
  - 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 as convolution kernels





Trick: Ensembles based on Heterogeneous Methods



Combination	Corpus	P	R	F
Single best				
APG	AIMed	59.9	53.6	56.2
APG	BioInfer	60.2	61.3	60.7
kBSPS	HPRD50	60.0	**88.4**	70.2
APG	IEPA	66.6	82.6	73.1
kBSPS	LLL	69.9	95.9	79.3
APG+SL+kBSPS	AIMed	58.0	**61.1**	**58.9**
	BioInfer	60.3	66.4	**63.0**
	HPRD50	67.6	76.9	**71.4**
	IEPA	68.6	**85.3**	**75.4**
	LLL	71.7	94.5	80.0

Are Confidence Values good Indicators? [Thomas et al. 2015]

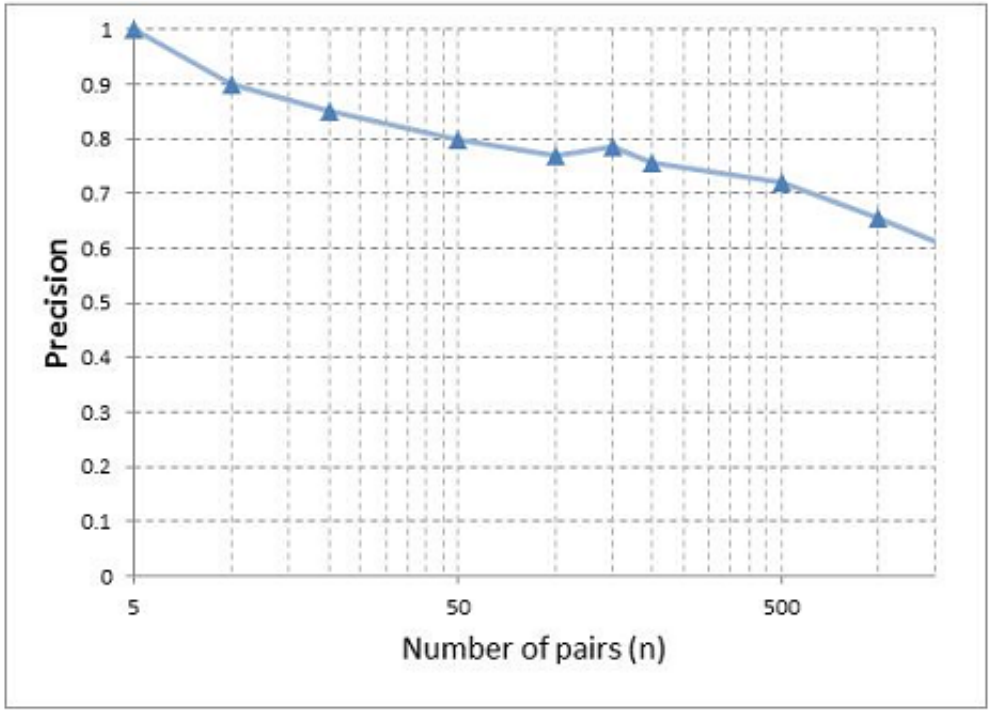
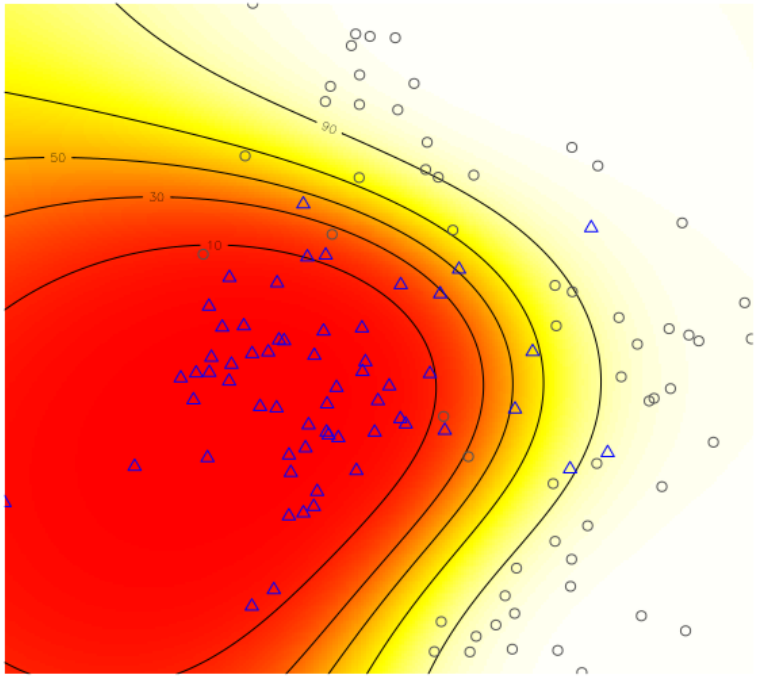
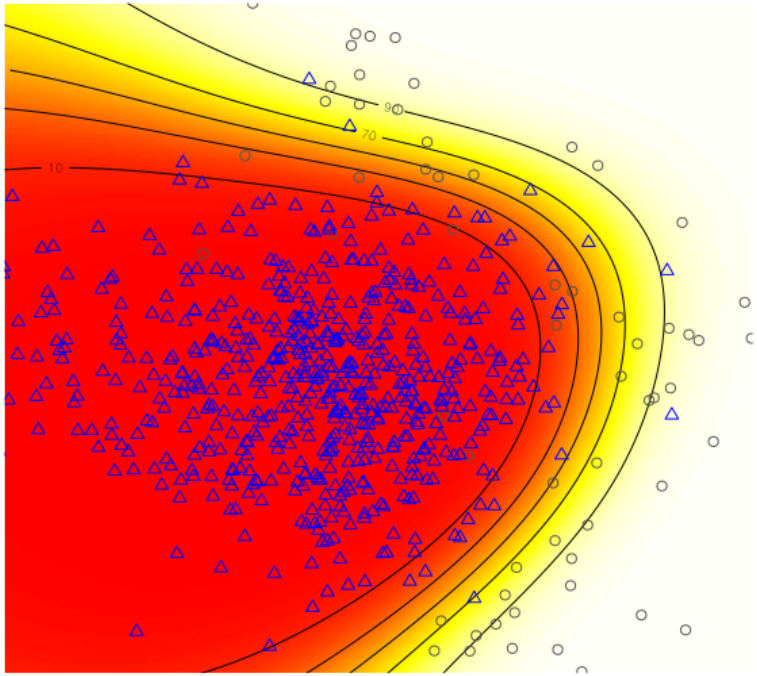


Fig. 4. Precision of our workflow for the n most confidently classified and manually curated sentences. Pairs already contained in a regulatory database are ignored (see Table 3).

Classifier tend to predict majority class

- Balanced/Unbalanced data set (same distribution) and learn a classifier





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

Literature

- Thomas, P. (2015). "Robust relationship extraction in the biomedical domain". Dissertation, Humboldt-Universität zu Berlin.
- Thomas, P., Durek, P., Solt, I., Klinger, B., Witzel, F., Schulthess, P., Mayer, Y., Tikk, D., Blüthgen, N. and Leser, U. (2015). "Computer-assisted curation of a human regulatory core network from the biological literature." *Bioinformatics 31(8): 1258-1266.*
- Tikk, D., Solt, I., Thomas, P. and Leser, U. (2013). "A Detailed Error Analysis of 13 Kernel Methods for Protein--Protein Interaction Extraction." *BMC Bioinformatics 14.*
- Tikk, D., Thomas, P., Palaga, P., Hakenberg, J. and Leser, U. (2010). "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature." *PLOS Computational Biology 6(7).*
- Hakenberg, J., Plake, C., Royer, L., Strobelt, H., Leser, U. and Schroeder, M. (2008). "Gene mention normalization and interaction extraction with context models and sentence motifs." *Genome Biol 9 Suppl 2: S14.*
- Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F. and Salakoski, T. (2008). "Comparative analysis of five protein-protein interaction corpora." *BMC Bioinformatics 9 Suppl 3: S6.*

Self-Assessment

- Give an upper bound on the accuracy of binary relationship extraction based on the accuracy of entity recognition
- How does co-occurrence-based RE work? Describe tricks to improve the expected performance. For each idea, describe the expected impact on precision and on recall.
- One problem of co-occurrence-based RE are expressions of the form "X is associated to A, B, C, and D". Imagine you had a method to detect such expressions. How could it be used to improve RE?
- Distant supervision for RE users automatically generated training data of unsure quality. Describe three ideas on how such data could be generated.