Statistical Parsing for Text Mining from Scientific Articles

Ted Briscoe

Computer Laboratory
University of Cambridge

November 30, 2004

Contents

Text Mining (TM)

Information Retrieval (IR)	finding relevant documents
Information Extraction (IE)	finding nuggets of information
Knowledge Discovery (KD)	discovering patterns of information

Textual Variation and TM

- 1) *The AntP protein represses BicD.*
- 1) *The AntP protein was found by Fujisaki et al.(1991) to repress BicD.*
- 1) *BicD is repressed by AntP.*
- 1) *BicD's repression by AntP is well-known.*
- 2) *It appears to physically interact with CLARP, a caspase-like molecule.*

IE/KD Subtasks

Noun phrases:	(NP *the_AT AntP_NN protein_NN1*)
Named Entity Classification:	(NP/gene *BiCD*)
Coreference:	(NP *It*) = (NP *The AntP protein*)
Relations:	repress(AntP, BiCD)
Modal Context:	appear(interact(AntP, CLARP))
Word Senses:	*drug/protein represses appetite/gene*

Grammatical Relations

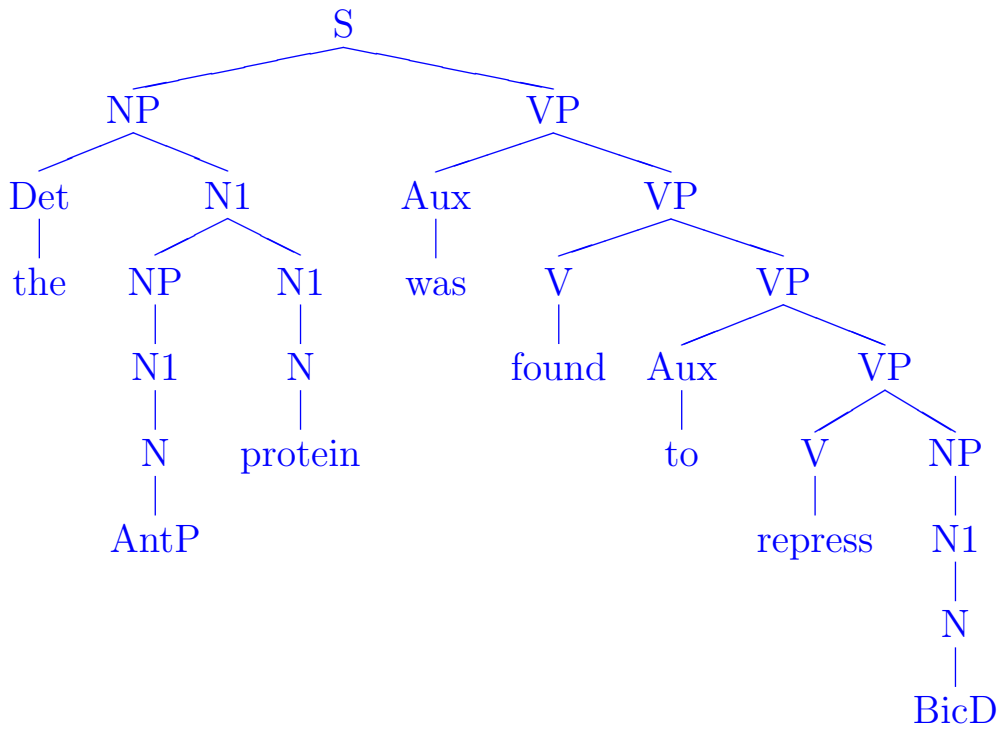
The AntP protein was found to repress BicD

(det protein the)
(nmod protein AntP)
(subj find protein obj)
(xcomp find repress)
(dobj repress BicD)
(subj repress protein)
(aux find be)

Dependency Graph (DAG) – close to Pred-Arg / F- Structure

Phrase Structure Tree

The AntP protein was found to repress BicD



Lexicalised Treebank Parsers

- ```
  (S (NP (NP Pierre Vinken) ,
         (NP (NP 61 years)
             (ADJP old))    ,)
      will (VP join
             (NP the board)
             (PP as
              (NP a nonexecutive director))
             (ADVP (NP Nov 29))))
  ```
- 10k+ CF rules conditioned on lexical items and structure (500k lexical parameters. $= +1\%$, Gildea, EMNLP01)
- Tree Recovery: approx 90% F-score
- Grammatical Relations: approx 70% F-score
  ```

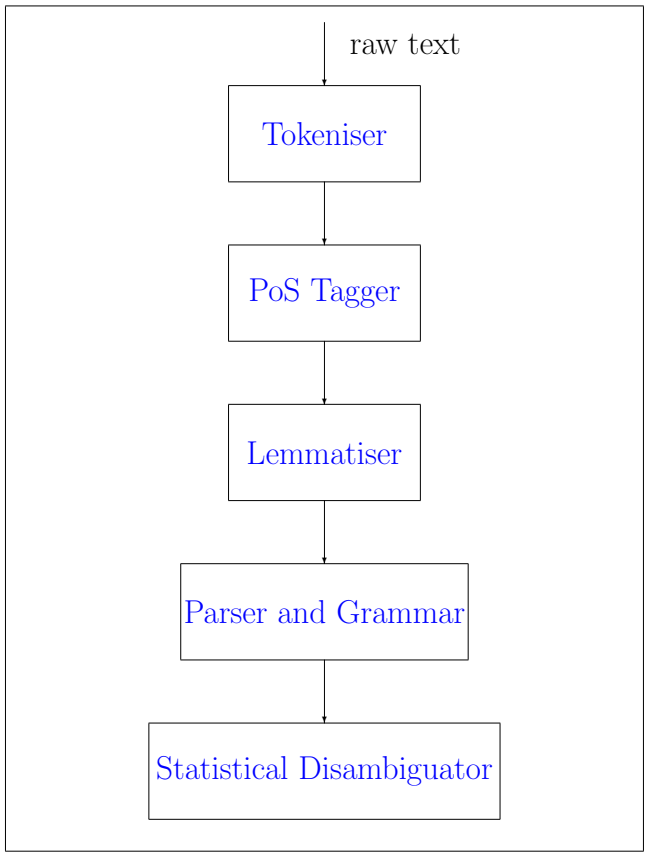
Lexicalised Treebank Parsers

- ```
  (S (NP (NP Pierre Vinken) ,
         (NP (NP 61 years)
             (ADJP old))    ,)
      will (VP join
             (NP the board)
             (PP as
              (NP a nonexecutive director))
             (ADVP (NP Nov 29))))
  ```
- 10k+ CF rules conditioned on lexical items and structure (500k lexical parameters. $= +1\%$, Gildea, EMNLP01)
- Tree Recovery: approx 90% F-score
- Grammatical Relations: approx 70% F-score
  ```

Lexicalised Treebank Parsers

- ```
  (S (NP (NP Pierre Vinken) ,
         (NP (NP 61 years)
             (ADJP old))    ,)
      will (VP join
             (NP the board)
             (PP as
              (NP a nonexecutive director))
             (ADVP (NP Nov 29))))
  ```
- 10k+ CF rules conditioned on lexical items and structure (500k lexical parameters. $= +1\%$, Gildea, EMNLP01)
- Tree Recovery: approx 90% F-score
- Grammatical Relations: approx 70% F-score
  ```

The RASP Pipeline



Tokenisation

- Sentence boundary detection: etc.
- Separation of punctuation: (Fred);
- Deterministic FST (in Flex/C)
- 200k words/sec, errors not evaluated

PoS and Punctuation Tagging

- About 150 extended PoS + punctuation tags: NN1
- 1st order HMM using FB algorithm in C (Elworthy, ANLP94)
- thresholded tag 'lattice' (4–10-fold decrease in tag error rate)
- 10k words/sec, approx. 0.3% error rate

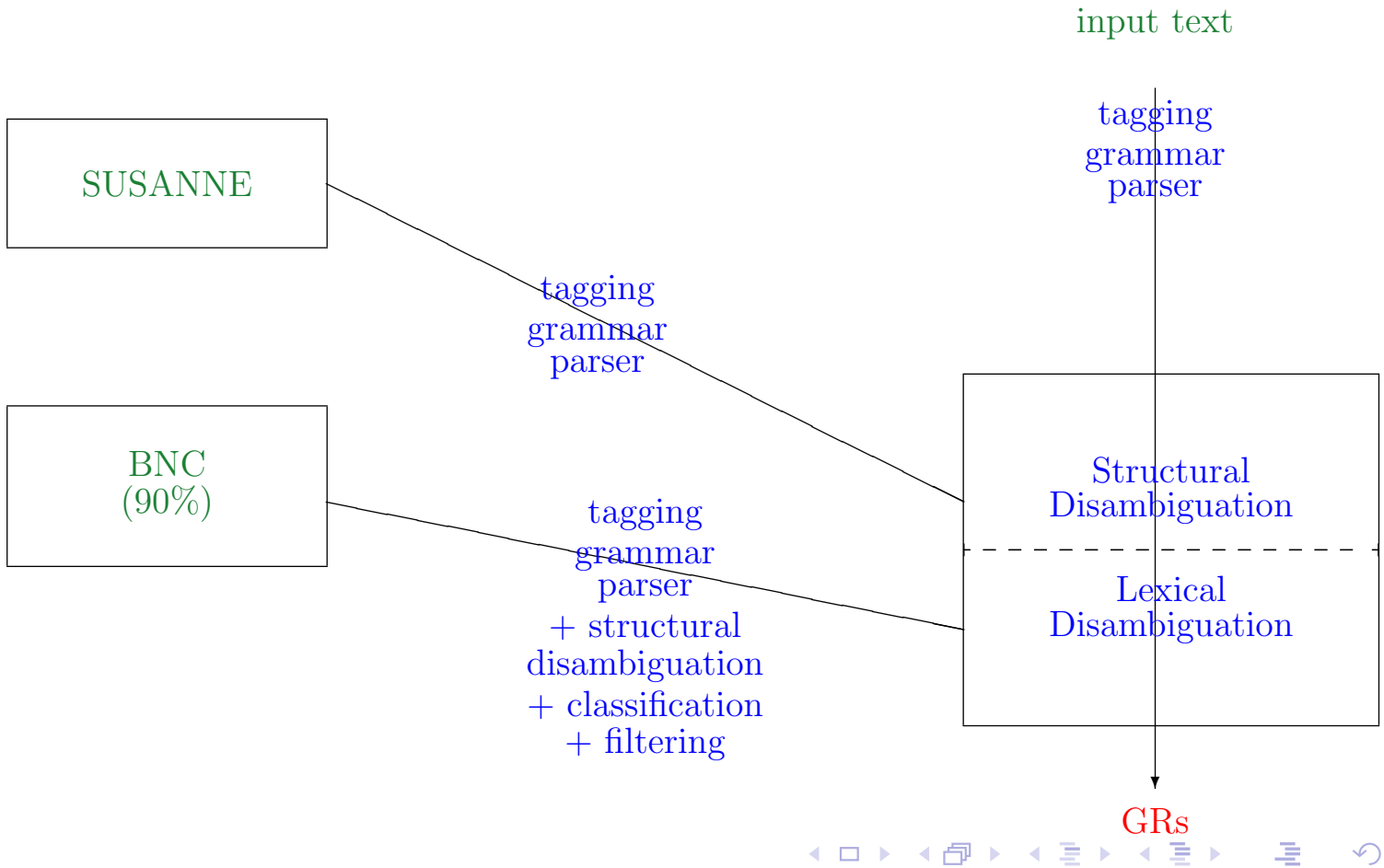
Lemmatisation

- 1400 FS rules (Minnen *et al.* JNLE01)
- Deterministic FST (in Flex/C)
- word/tag to lemma+affix/tag run+ing_VVG
- 200k words/sec, 0.007% error rate

Probabilistic Generalised LR Parser

- Manually-created, feature-based generic grammar (1.8K rules)
- Probabilities: shift/reduce conflicts, lookahead item (Briscoe and Carroll CL93)
- Partial-parse recovery computes shortest path in weighted graph-structured stack (Kiefer *et al.* ACL99)
- Parse forest construction is (experimentally) quadratic in sentence length (av. 100 tokens/sec)

Parse Ranking



Ambiguity Resolution

The young man the boats

```
The AT:1
young JJ:0.999544 NN1:0.000456463
man NN1:0.999839 VV0:0.000161307
the AT:1
boats NN2:1
```

```
(|ncsubj| |man:3_VV0| |young:2_NN1| _)
(|dobj| |man:3_VV0| |boat+s:5_NN2| _)
(|det |young:2_NN1| |The:1_AT|)
(|det |boat+s:5_NN2| |the:4_AT|)
```

Ambiguity Resolution

The young man the boats

```
The AT:1
young JJ:0.999544 NN1:0.000456463
man NN1:0.999839 VV0:0.000161307
the AT:1
boats NN2:1
```

```
(|ncsubj| |man:3_VV0| |young:2_NN1| _)
(|dobj| |man:3_VV0| |boat+s:5_NN2| _)
(|det |young:2_NN1| |The:1_AT|)
(|det |boat+s:5_NN2| |the:4_AT|)
```

Weighted Grammatical Relations

Peter reads every paper on markup

```
1.0 (|ncsubj| |reads:2_VVZ| |Peter:1_NP1| _)
0.7 (|nmod| |on:5_II| |paper:4_NN1| |markup:6_NN|)
0.3 (|nmod| |on:5_II| |reads:2_VVZ| |markup:6_NN|)
```

GR Evaluation Experiment

Maximum Parses (n)	Precision (%)	Recall (%)	F-score
1	76.25	76.77	76.51
2	80.15	73.30	76.57
5	84.94	67.03	74.93
10	86.73	62.47	72.63
100	89.59	51.45	65.36
1000	90.24	46.08	61.00
unlimited	90.40	45.21	60.27

A Curated database

- 10 curators download and read papers
- Watch list of 35 journals

- Gene proformas (IE templates, 34 fields)
  ! G1a. Gene symbol to use in database *a :
  ! G1b. Gene symbol used in paper (if different) *i :
  ! G4b. Other synonym(s) for gene symbol *i :
  ! G20b. G1a wildtype expression in wildtype analysed? NSC :
  ! G18. Gene(s) stated to interact genetically with G1a *p :
  ...

- Automated integration of proformas with the database

A Curated database

- 10 curators download and read papers
- Watch list of 35 journals

- Gene proformas (IE templates, 34 fields)
  ! G1a. Gene symbol to use in database *a :
  ! G1b. Gene symbol used in paper (if different) *i :
  ! G4b. Other synonym(s) for gene symbol *i :
  ! G20b. G1a wildtype expression in wildtype analysed? NSC :
  ! G18. Gene(s) stated to interact genetically with G1a *p :
  ...

- Automated integration of proformas with the database

A Curated database

- 10 curators download and read papers
- Watch list of 35 journals

- Gene proformas (IE templates, 34 fields)
  ! G1a. Gene symbol to use in database *a :
  ! G1b. Gene symbol used in paper (if different) *i :
  ! G4b. Other synonym(s) for gene symbol *i :
  ! G20b. G1a wildtype expression in wildtype analysed? NSC :
  ! G18. Gene(s) stated to interact genetically with G1a *p :
  ...

- Automated integration of proformas with the database

Analysis of Curated Papers Archive

1. 5k+ articles in PDF or XML format
2. PDF → XML, mark-up of sections, etc
3. Parse document with RASP
4. Link text passages to gene proformas
5. Interactive construction of proformas

Analysis of Curated Papers Archive

1. 5k+ articles in PDF or XML format
2. PDF → XML, mark-up of sections, etc
3. Parse document with RASP
4. Link text passages to gene proformas
5. Interactive construction of proformas

Analysis of Curated Papers Archive

1. 5k+ articles in PDF or XML format
2. PDF → XML, mark-up of sections, etc
3. Parse document with RASP
4. Link text passages to gene proformas
5. Interactive construction of proformas

Analysis of Curated Papers Archive

1. 5k+ articles in PDF or XML format
2. PDF → XML, mark-up of sections, etc
3. Parse document with RASP
4. Link text passages to gene proformas
5. Interactive construction of proformas

Analysis of Curated Papers Archive

1. 5k+ articles in PDF or XML format
2. PDF $\rightarrow$ XML, mark-up of sections, etc
3. Parse document with RASP
4. Link text passages to gene proformas
5. Interactive construction of proformas

FlyBase as a Dictionary

- 18k Genes, 75k Gene names / acronyms
- Overlap with general English: AN, BUT, CAN, MAD, spliced
- Spelling variation: FAS-III, fas III
- Synonymy (2+ terms / 1 gene): PHM, dPHM
- Homonymy (1 term / 2+ entities): iab

Integrating Named Entity Recognition with RASP

- *The interleukin-2 (IL-2) promoter ...*
- (det promoter The)
  (nmod brack IL-2 promoter)
  (nmod _ Interleukin-2 promoter)
- *The c-rel and v-rel (proto) oncogenes ...*
-

a)
(det oncogene+s The)
(nmod brack proto oncogene+s)
(nmod _ c-rel oncogene+s)
(ncmod _ v-rel oncogene+s)
(conj and c-rel v-rel)

b)
(det oncogene+s The)
(nmod brack proto oncogene+s)
(det c-rel The)
(nmod _ v-rel oncogene+s)
(conj and c-rel oncogene+s)

Integrating Named Entity Recognition with RASP

- *The interleukin-2 (IL-2) promoter ...*
- (det promoter The)
  (nmod brack IL-2 promoter)
  (nmod _ Interleukin-2 promoter)
- *The c-rel and v-rel (proto) oncogenes ...*
-

a)
(det oncogene+s The)
(nmod brack proto oncogene+s)
(nmod _ c-rel oncogene+s)
(ncmod _ v-rel oncogene+s)
(conj and c-rel v-rel)

b)
(det oncogene+s The)
(nmod brack proto oncogene+s)
(det c-rel The)
(nmod _ v-rel oncogene+s)
(conj and c-rel oncogene+s)

Integrating Named Entity Recognition with RASP

- *The interleukin-2 (IL-2) promoter ...*
- (det promoter The)
  (nmod brack IL-2 promoter)
  (nmod _ Interleukin-2 promoter)
- *The c-rel and v-rel (proto) oncogenes ...*
-

  a)
  (det oncogene+s The)
  (nmod brack proto oncogene+s)
  (nmod _ c-rel oncogene+s)
  (ncmod _ v-rel oncogene+s)
  (conj and c-rel v-rel)

  b)
  (det oncogene+s The)
  (nmod brack proto oncogene+s)
  (det c-rel The)
  (nmod _ v-rel oncogene+s)
  (conj and c-rel oncogene+s)

Domain Predicate Resolution

1. Parse papers archive with generic RASP system
2. Statistically induce words' properties from parse contexts

- (S *AntP represses_Vtrans BicD*) $\rightsquigarrow$ repress(protein, gene)
- $\rightsquigarrow$ *The represssion of BicD by AntP*

Domain Predicate Resolution

1. Parse papers archive with generic RASP system
2. Statistically induce words' properties from parse contexts

- (S *AntP represses_Vtrans BicD*) $\rightsquigarrow$ repress(protein, gene)
- $\rightsquigarrow$ *The represssion of BicD by AntP*

Domain Predicate Resolution

1. Parse papers archive with generic RASP system
2. Statistically induce words' properties from parse contexts

- (S *AntP represses_Vtrans BicD*) $\rightsquigarrow$ repress(protein, gene)
- $\rightsquigarrow$ *The represssion of BicD by AntP*

Domain Predicate Resolution

1. Parse papers archive with generic RASP system
2. Statistically induce words' properties from parse contexts

- (S *AntP represses_Vtrans BicD*) $\rightsquigarrow$ repress(protein, gene)
- $\rightsquigarrow$ *The represssion of BicD by AntP*

TM: IE

1. GRs support recovery of information nuggets:
   S^n:(subj repress protein _) $\wedge$ (dobj repress BicD _)

2. in modal contexts:
   S^{n+1}:(xcomp to find repress)

3. and anaphora resolution:
   S^{n+1}:(subj appear it) $\rightsquigarrow$ S^n:(subj find protein obj)

TM: IE

1. GRs support recovery of information nuggets:
   S^n:(subj repress protein _) $\wedge$ (dobj repress BicD _)

2. in modal contexts:
   S^{n+1}:(xcomp to find repress)

3. and anaphora resolution:
   S^{n+1}:(subj appear it) $\rightsquigarrow$ S^n:(subj find protein obj)

TM: IE

1. GRs support recovery of information nuggets:
   S^n:(subj repress protein _) $\wedge$ (dobj repress BicD _)
2. in modal contexts:
   S^{n+1}:(xcomp to find repress)
3. and anaphora resolution:
   S^{n+1}:(subj appear it) $\rightsquigarrow$ S^n:(subj find protein obj)

TM: IR/KD: Passage Retrieval and Classification

- More precise queries:
  AntP, *BicD*, *repress*, ... $\rightsquigarrow$ repress(AntP, BicD)
- Gene expression passages:
  AntP, *BicD*, *express*, ... $\rightsquigarrow$ express(gene, protein)

TM: IR/KD: Passage Retrieval and Classification

- More precise queries:
  AntP, *BicD*, *repress*, ... $\rightsquigarrow$ repress(AntP, BicD)
- Gene expression passages:
  AntP, *BicD*, *express*, ... $\rightsquigarrow$ express(gene, protein)

The Semantic Web Context

- Domain Resources: migration to XML $\rightsquigarrow$ RDF, OWL Lite e.g. Gene Sequence Ontology
- IE Customisation Standards: Lexicon and Ontology APIs, etc
- IE as Layered Annotation: XML (standoff) pipeline $\rightsquigarrow$ document metadata (RDF)
- RDF Metadata: semantic search and flexible integration with developing domain resources still indexed to text (passages)
- RDF-based Curation: gene proforma + provenance, textual evidence, links to ontologies, database and other literature
- SW Systems: MnM, KIM, etc: integrating GATE (IE as XML annot.), Sesame (RDF repository), and Lucene (IR engine)

The Semantic Web Context

- Domain Resources: migration to XML $\rightsquigarrow$ RDF, OWL Lite e.g. Gene Sequence Ontology
- IE Customisation Standards: Lexicon and Ontology APIs, etc
- IE as Layered Annotation: XML (standoff) pipeline $\rightsquigarrow$ document metadata (RDF)
- RDF Metadata: semantic search and flexible integration with developing domain resources still indexed to text (passages)
- RDF-based Curation: gene proforma + provenance, textual evidence, links to ontologies, database and other literature
- SW Systems: MnM, KIM, etc: integrating GATE (IE as XML annot.), Sesame (RDF repository), and Lucene (IR engine)

The Semantic Web Context

- Domain Resources: migration to XML $\rightsquigarrow$ RDF, OWL Lite e.g. Gene Sequence Ontology
- IE Customisation Standards: Lexicon and Ontology APIs, etc
- IE as Layered Annotation: XML (standoff) pipeline $\rightsquigarrow$ document metadata (RDF)
- RDF Metadata: semantic search and flexible integration with developing domain resources still indexed to text (passages)
- RDF-based Curation: gene proforma + provenance, textual evidence, links to ontologies, database and other literature
- SW Systems: MnM, KIM, etc: integrating GATE (IE as XML annot.), Sesame (RDF repository), and Lucene (IR engine)

The Semantic Web Context

- Domain Resources: migration to XML $\rightsquigarrow$ RDF, OWL Lite e.g. Gene Sequence Ontology
- IE Customisation Standards: Lexicon and Ontology APIs, etc
- IE as Layered Annotation: XML (standoff) pipeline $\rightsquigarrow$ document metadata (RDF)
- RDF Metadata: semantic search and flexible integration with developing domain resources still indexed to text (passages)
- RDF-based Curation: gene proforma + provenance, textual evidence, links to ontologies, database and other literature
- SW Systems: MnM, KIM, etc: integrating GATE (IE as XML annot.), Sesame (RDF repository), and Lucene (IR engine)

The Semantic Web Context

- Domain Resources: migration to XML $\rightsquigarrow$ RDF, OWL Lite e.g. Gene Sequence Ontology
- IE Customisation Standards: Lexicon and Ontology APIs, etc
- IE as Layered Annotation: XML (standoff) pipeline $\rightsquigarrow$ document metadata (RDF)
- RDF Metadata: semantic search and flexible integration with developing domain resources still indexed to text (passages)
- RDF-based Curation: gene proforma + provenance, textual evidence, links to ontologies, database and other literature
- SW Systems: MnM, KIM, etc: integrating GATE (IE as XML annot.), Sesame (RDF repository), and Lucene (IR engine)

The Semantic Web Context

- Domain Resources: migration to XML $\rightsquigarrow$ RDF, OWL Lite e.g. Gene Sequence Ontology
- IE Customisation Standards: Lexicon and Ontology APIs, etc
- IE as Layered Annotation: XML (standoff) pipeline $\rightsquigarrow$ document metadata (RDF)
- RDF Metadata: semantic search and flexible integration with developing domain resources still indexed to text (passages)
- RDF-based Curation: gene proforma + provenance, textual evidence, links to ontologies, database and other literature
- SW Systems: MnM, KIM, etc: integrating GATE (IE as XML annot.), Sesame (RDF repository), and Lucene (IR engine)

Conclusions

- **TM = IR + IE + KD** in **SW**
- TM expensive to develop, apart from IR
- Adaptive IE deeper statistical analysis, more generic rules, (weakly-supervised) domain customisation
- Semantic Web infrastructure for standardising domain customisation using extant domain resources
- Web Services e.g. TM as a component of microarray data interpretation

Conclusions

- **TM = IR + IE + KD** in **SW**
- TM expensive to develop, apart from IR
- Adaptive IE deeper statistical analysis, more generic rules, (weakly-supervised) domain customisation
- Semantic Web infrastructure for standardising domain customisation using extant domain resources
- Web Services e.g. TM as a component of microarray data interpretation

Conclusions

- **TM = IR + IE + KD** in **SW**
- TM expensive to develop, apart from IR
- Adaptive IE deeper statistical analysis, more generic rules, (weakly-supervised) domain customisation
- Semantic Web infrastructure for standardising domain customisation using extant domain resources
- Web Services e.g. TM as a component of microarray data interpretation

Conclusions

- **TM = IR + IE + KD** in **SW**
- TM expensive to develop, apart from IR
- Adaptive IE deeper statistical analysis, more generic rules, (weakly-supervised) domain customisation
- Semantic Web infrastructure for standardising domain customisation using extant domain resources
- Web Services e.g. TM as a component of microarray data interpretation

Conclusions

- **TM = IR + IE + KD** in **SW**
- TM expensive to develop, apart from IR
- Adaptive IE deeper statistical analysis, more generic rules, (weakly-supervised) domain customisation
- Semantic Web infrastructure for standardising domain customisation using extant domain resources
- Web Services e.g. TM as a component of microarray data interpretation