The annotation conundrum

Mark Liberman

University of Pennsylvania
myl@cis.upenn.edu

The setting

- There are many kinds of linguistic annotation:

  Phonetics, prosody, P.O.S., trees, word senses, co-reference, propositions, etc.

- This talk focuses on two specific, practical categories of annotation
  - "entities" : textual references to things of a given type
    - people, places, organizations, genes, diseases …
    - may be normalized as a second step

      "Myanmar" = "Burma"
      "5/26/2008" = "26/05/2008" = "May 26, 2008" = etc.
  - "relations" among entities
    - <person> employed by <organization>
    - <genomic variation> associated with <disease state>

- Recipe for an entity (or relation) tagger:
  - Humans tag a training set with typed entities (& relations)
  - Apply machine learning, and hope for F = 0.7 to 0.9
  - This is an active area for machine-learning research

- Good entity and relation taggers have many applications

Entity problems in MT

昨天下午，当记者乘坐的东航MU5413航班抵达四川成都“双流”机场时，迎接记者的就是青川发生6.4级余震。

Yesterday afternoon, as a reporter by the China Eastern flight MU5413 arrived in Chengdu, Sichuan "Double" at the airport, greeted the news is the Green-6.4 aftershock occurred.

双流 Shuāng liú	Shuangliu
双 shuāng	two; double; pair; both
流 liú	to flow; to spread; to circulate; to move
机场 jī chǎng	airport
青川 Qīng chuān	Qingchuan (place in Sichuan)
青 qīng	green (blue, black)
川 chuān	river; creek; plain; an area of level country

The problem

- “Natural annotation” is inconsistent

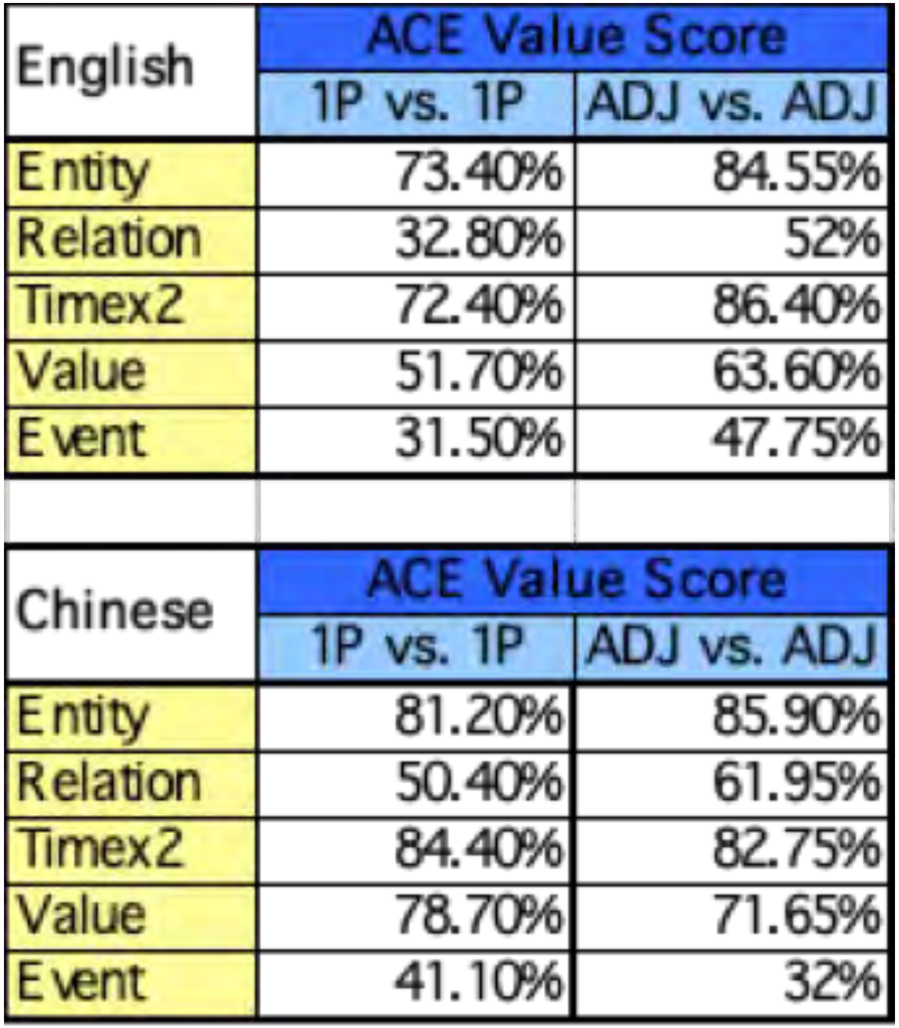
  Give annotators a few examples (or a simple definition), turn them loose, and you get:

  - poor agreement for entities (often F=0.5 or worse)
  - worse for normalized entities
  - worse yet for relations
- Why?
  - Human generalization from examples is variable
  - Human application of principles is variable
  - NL context raises many hard questions:
    … treatment of modifiers, metonymy, hypo- and hypernyms, descriptions, recursion, irrealis contexts, referential vagueness, etc.
- As a result
  - The “gold standard” is not naturally very golden
  - The resulting machine learning metrics are noisy
- And F-score of 0.3-0.5 is not an attractive goal!

The traditional solution

- Iterative refinement of guidelines
  1. Try some annotation
  2. Compare and contrast
  3. Adjudicate and generalize
  4. Go back to 1 and repeat throughout project
     (or at least until inter-annotator agreement is adequate)
- Convergence is usually slow
- Result: a complex accretion of “common law”
  - Slow to develop and hard to learn
  - More consistent than “natural annotation”
    - But fit to applications (including theories) is unclear
  - Complexity may re-create inconsistency
    new types and sub-types → ambiguity, confusion

ACE 2005 (in)consistency



English	ACE Value Score	
	1P vs. 1P	ADJ vs. ADJ
Entity	73.40%	84.55%
Relation	32.80%	52%
Timex2	72.40%	86.40%
Value	51.70%	63.60%
Event	31.50%	47.75%

Chinese	ACE Value Score	
	1P vs. 1P	ADJ vs. ADJ
Entity	81.20%	85.90%
Relation	50.40%	61.95%
Timex2	84.40%	82.75%
Value	78.70%	71.65%
Event	41.10%	32%

- ***1P vs. 1P***
  independent first passes by junior annotator, no QC
- ***ADJ vs. ADJ***
  output of two parallel, independent dual first pass annotations are adjudicated by two independent senior annotators

Iterative improvement

From ACE 2005 (Ralph Weischedel):

Repeat until criteria met or until time has expired:

1. Analyze performance of previous task & guidelines
   - Scores, confusion matrices, etc.
2. Hypothesize & implement changes to tasks/guidelines
3. Update infrastructure as needed
   - DTD, annotation tool, and scorer
4. Annotate texts
5. Evaluate inter-annotator agreement

ACE as NLP judiciary

150 complex rules

- Plus Wiki
- Plus Listserv

Rules, Notes, Fiats and Exceptions		
Task	**#Pages**	**#Rules**
Entity	34	20
Value	10	5
TIMEX2	75	50
Relations	36	25
Events	77	50
Total	**232**	**150**

Example Decision Rule (Event p33)

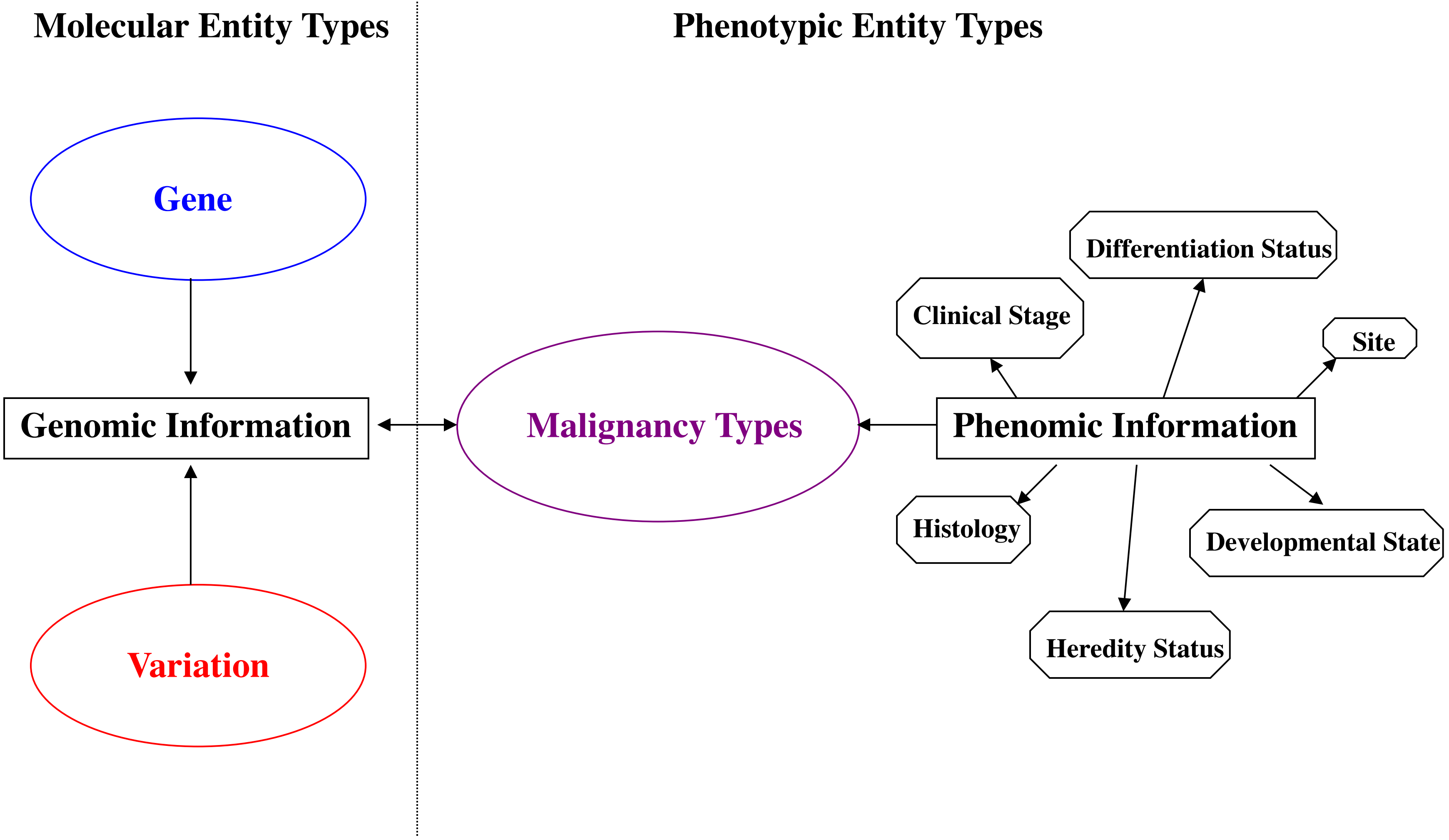
> ***Note:*** *For Events that where a single common trigger is ambiguous between the types* LIFE *(i.e.* INJURE *and* DIE) *and* CONFLICT *(i.e.* ATTACK)*, we will only annotate the Event as a* LIFE *Event in case the relevant resulting state is clearly indicated by the construction.*
>
> *The above rule will not apply when there are independent triggers.*

BioIE case law

Guidelines for oncology tagging

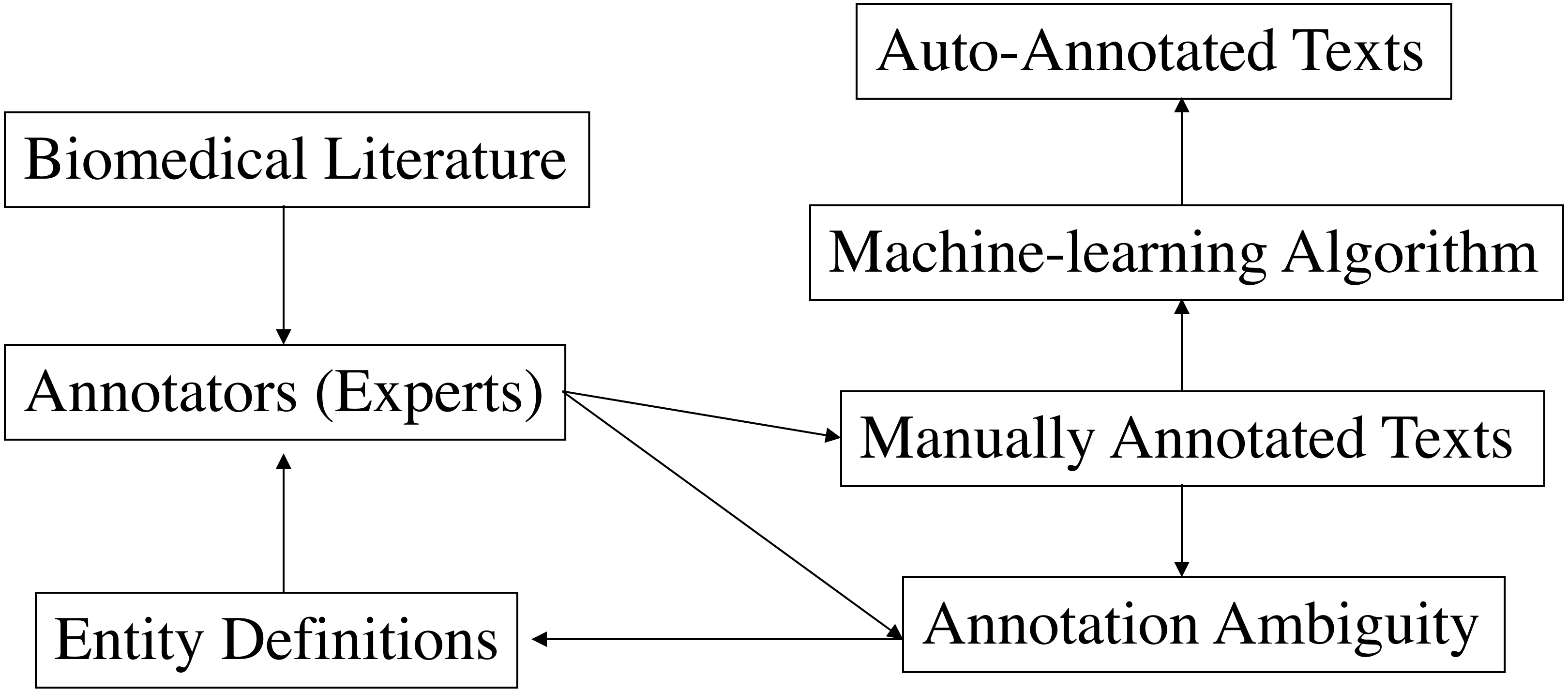
These were developed under the guidance
of Yang Jin (then a neuroscience graduate student
interested in the relationship between
genomic variations and neuroblastoma)
and his advisor, Dr. Pete White.

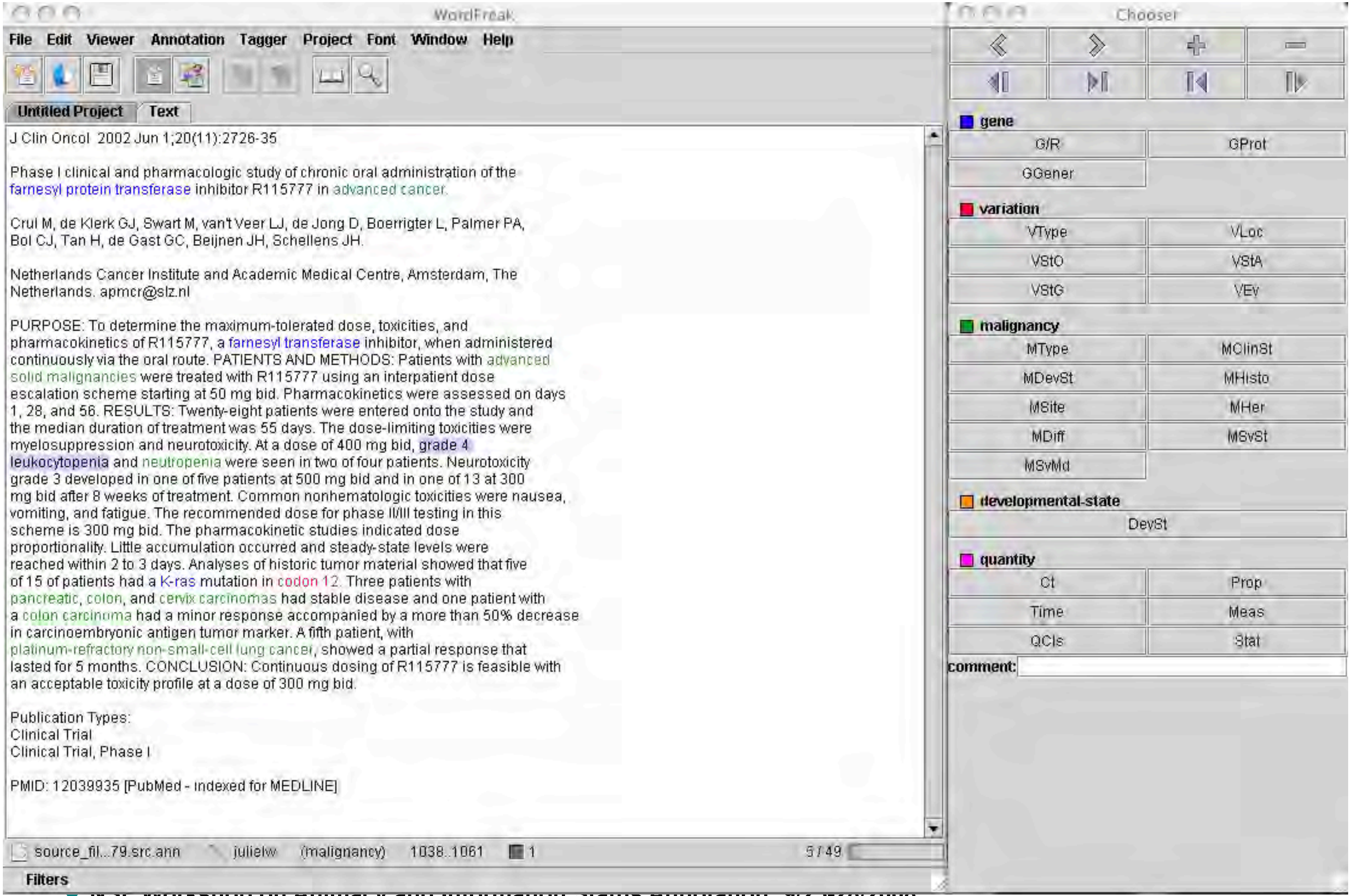
The result was a set of excellent taggers,
but the process was long and complex.



Genomic Variation associated with Malignancy

Flow Chart for Manual Annotation Process





J Clin Oncol 2002 Jun 1;20(11):2726-35

Phase I clinical and pharmacologic study of chronic oral administration of the farnesyl protein transferase inhibitor R115777 in advanced cancer.

Crul M, de Klerk GJ, Swart M, van't Veer LJ, de Jong D, Boerrigter L, Palmer PA, Bol CJ, Tan H, de Gast GC, Beijnen JH, Schellens JH.

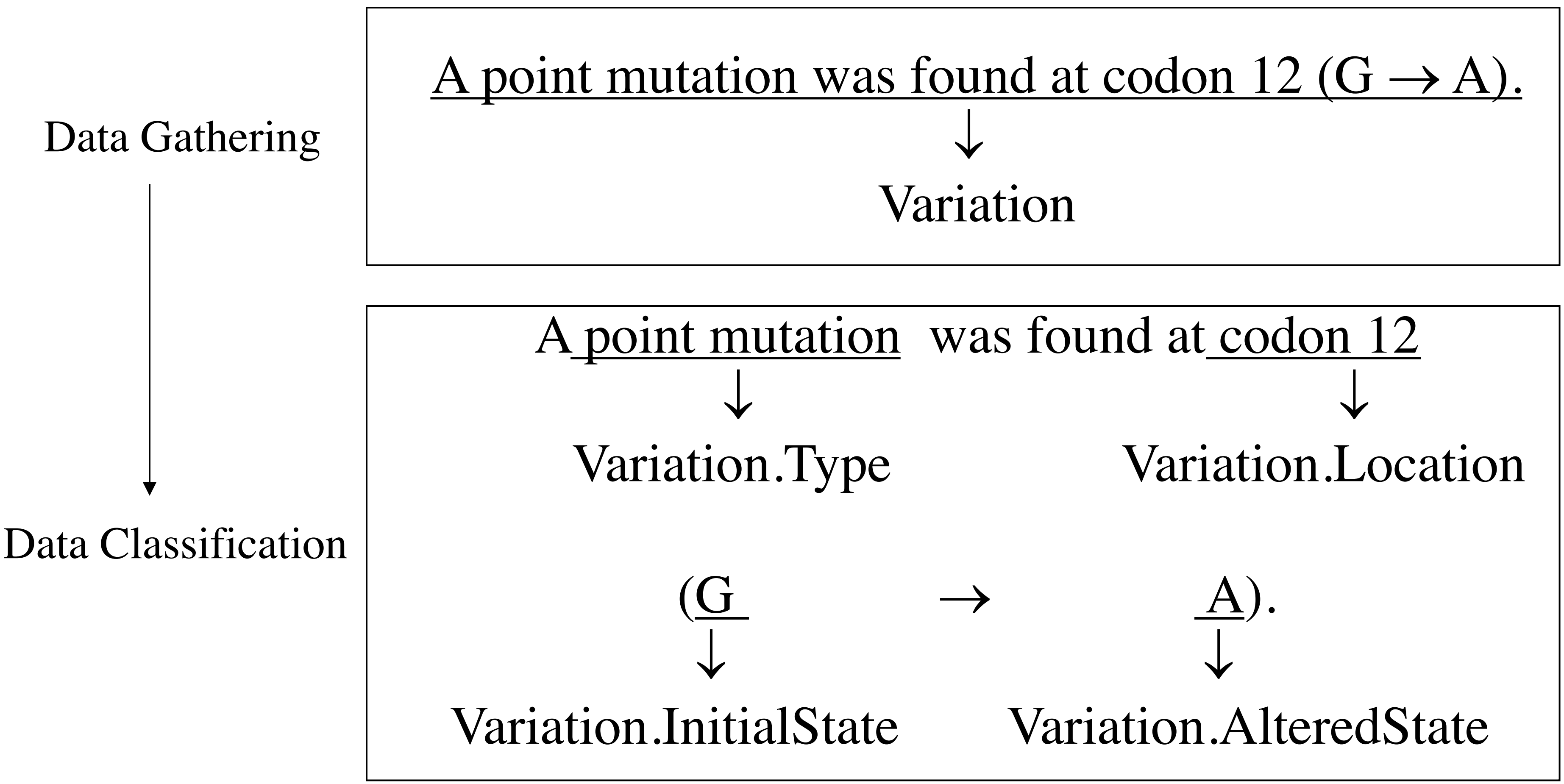
Netherlands Cancer Institute and Academic Medical Centre, Amsterdam, The Netherlands. apmcr@slz.nl

PURPOSE: To determine the maximum-tolerated dose, toxicities, and pharmacokinetics of R115777, a farnesyl transferase inhibitor, when administered continuously via the oral route. PATIENTS AND METHODS: Patients with advanced solid malignancies were treated with R115777 using an interpatient dose escalation scheme starting at 50 mg bid. Pharmacokinetics were assessed on days 1, 28, and 56. RESULTS: Twenty-eight patients were entered onto the study and the median duration of treatment was 55 days. The dose-limiting toxicities were myelosuppression and neurotoxicity. At a dose of 400 mg bid, grade 4 leukocytopenia and neutropenia were seen in two of four patients. Neurotoxicity grade 3 developed in one of five patients at 500 mg bid and in one of 13 at 300 mg bid after 8 weeks of treatment. Common nonhematologic toxicities were nausea, vomiting, and fatigue. The recommended dose for phase II/III testing in this scheme is 300 mg bid. The pharmacokinetic studies indicated dose proportionality. Little accumulation occurred and steady-state levels were reached within 2 to 3 days. Analyses of historic tumor material showed that five of 15 of patients had a K-ras mutation in codon 12. Three patients with pancreatic, colon, and cervix carcinomas had stable disease and one patient with a colon carcinoma had a minor response accompanied by a more than 50% decrease in carcinoembryonic antigen tumor marker. A fifth patient, with platinum-refractory non-small-cell lung cancer, showed a partial response that lasted for 5 months. CONCLUSION: Continuous dosing of R115777 is feasible with an acceptable toxicity profile at a dose of 300 mg bid.

Publication Types:
Clinical Trial
Clinical Trial, Phase I

PMID: 12039935 [PubMed - indexed for MEDLINE]

Defining biomedical entities



Data Gathering

A point mutation was found at codon 12 (G → A).

↓

Variation

Data Classification

A point mutation was found at codon 12

↓ Variation.Type

↓ Variation.Location

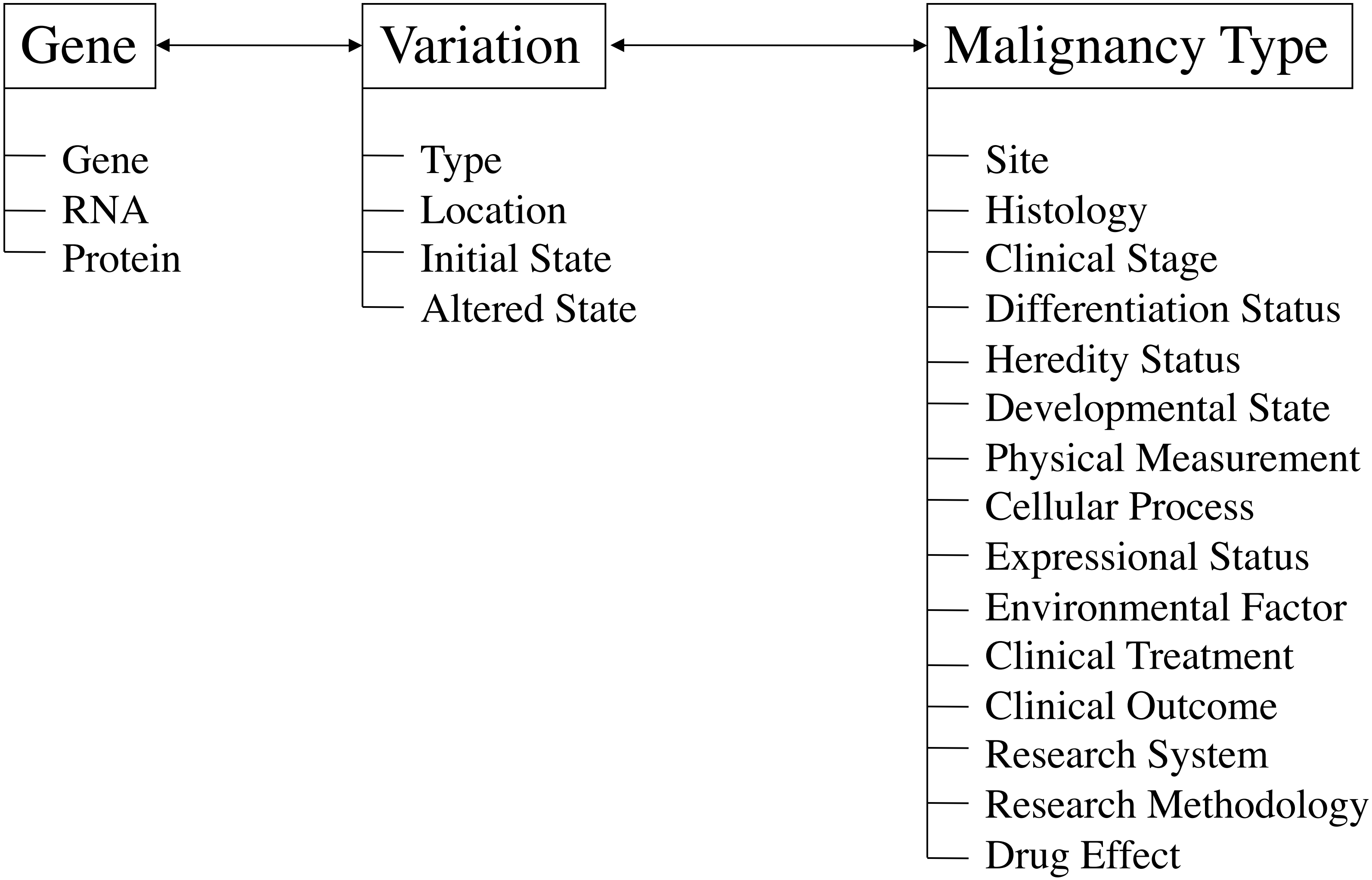
(G → A).

↓ Variation.InitialState

↓ Variation.AlteredState

Defining biomedical entities

- Conceptual issues
  - Sub-classification of entities
  - Levels of specificity
    - MAPK10, MAPK, protein kinase, gene
    - squamous cell lung carcinoma, lung carcinoma, carcinoma, cancer
  - Conceptual overlaps between entities *(e.g. symptom vs. disease)*
- Linguistic issues
  - Text boundary issues (The *K-ras* gene)
  - Co-reference (this gene, it, they)
  - Structural overlap -- entity within entity
    - squamous cell lung carcinoma
    - MAP kinase kinase kinase
  - Discontinuous mentions (*N*- and *K-ras*)



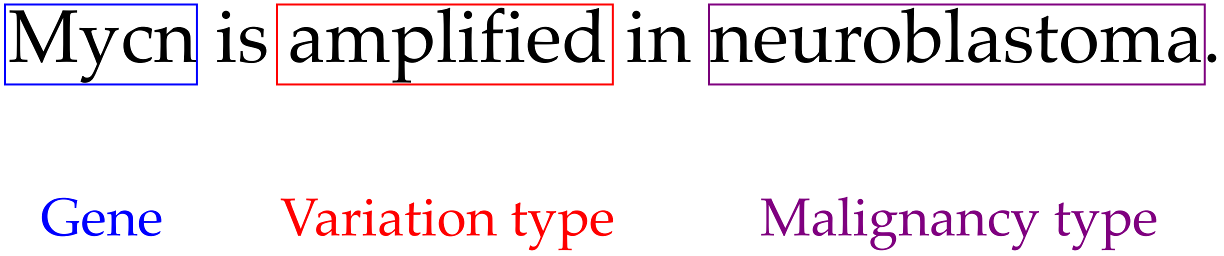
Gene ↔ Variation ↔ Malignancy Type

Gene
- Gene
- RNA
- Protein

Variation
- Type
- Location
- Initial State
- Altered State

Malignancy Type
- Site
- Histology
- Clinical Stage
- Differentiation Status
- Heredity Status
- Developmental State
- Physical Measurement
- Cellular Process
- Expressional Status
- Environmental Factor
- Clinical Treatment
- Clinical Outcome
- Research System
- Research Methodology
- Drug Effect

Named Entity Extractors



Mycn is amplified in neuroblastoma.

Gene

Variation type

Malignancy type

Automated Extractor Development

- Training and testing data
  - 1442 cancer-focused MEDLINE abstracts
  - 70% for training, 30% for testing
- Machine-learning algorithm
  - Conditional Random Fields (CRFs)
  - Sets of Features
    - Orthographic features (capitalization, punctuation, digit/number/alpha-numeric/symbol);
    - Character-N-grams (N=2,3,4);
    - Prefix/Suffix: (*oma);
    - Nearby words;
    - Domain-specific lexicon (NCI neoplasm list).

Extractor Performance

Entity	Precision	Recall
Gene	0.864	0.787
Variation Type	0.8556	0.7990
Location	0.8695	0.7722
State-Initial	0.8430	0.8286
State-Sub	0.8035	0.7809
Overall	0.8541	0.7870
Malignancy type	0.8456	0.8218
Clinical Stage	0.8493	0.6492
Site	0.8005	0.6555
Histology	0.8310	0.7774
Developmental State	0.8438	0.7500

- Precision: (true positives)/(true positives + false positives)
- Recall: (true positives)/(true positives + false negatives)

Normal text
Malignancies

PMID: 15316311

Morphologic and molecular characterization of *renal cell carcinoma* in children and young adults

A new WHO classification of *renal cell carcinoma* has been introduced in 2004. This classification includes the recently described *renal cell carcinomas* with the ASPL-TFE3 gene fusion and *carcinomas* with a PRCC-TFE3 gene fusion. Collectively, these tumors have been termed Xp11.2 or TFE3 *translocation carcinomas*, which primarily occur in children and young adults. To further study the characteristics of *renal cell carcinoma* in young patients and to determine their genetic background, 41 *renal cell carcinomas* of patients younger than 22 years were morphologically and genetically characterized Loss of heterozygosity analysis of the von Hippel-Lindau gene region and screening for VHL gene mutations by direct sequencing were performed in 20 tumors. TFE3 protein overexpression, which correlates with the presence of a TFE3 gene fusion, was assessed by immunohistochemistry. Applying the new WHO classification for *renal cell carcinoma*, there were 6 clear cell (15%), 9 papillary (22%), 2 chromophobe, and 2 collecting duct *carcinomas*. Eight *carcinomas* showed translocation carcinoma morphology (20%). One *carcinoma* occurred 4 years after a *neuroblastoma*. Thirteen tumors could not be assigned to types specified by the new WHO classification: 10 were grouped as unclassified (24%), including a unique *renal cell carcinoma* with prominently vacuolated cytoplasm and WT1 expression. Three *carcinomas* occurred in combination with *nephroblastoma*. Molecular analysis revealed deletions at 3p25-26 in one *translocation carcinoma*, one *chromophobe renal cel carcinoma*, and one *papillary renal cel carcinoma*. There were no VHL mutations. Nuclear TFE3 overexpression was detected in 6 *renal cell carcinomas*, all of which showed areas with voluminous cytoplasm and foci of papillary architecture, consistent with a *translocation carcinoma* phenotype. The large proportion of TFE3 "translocation "*carcinomas* and "unclassified" *carcinomas* in the first two decades of life demonstrates that *renal cell carcinomas* in young patients contain genetically and phenotypically distinct tumors with further potential for novel *renal cell carcinoma* subtypes. The far lower frequency of *clear cell carcinomas* and VHL alterations compared with adults suggests that *renal cell carcinomas* in young patients have a unique genetic background.

CRF-based Extractor vs. Pattern Matcher

- The testing corpus
  - 39 manually annotated MEDLINE abstracts selected
  - 202 malignancy type mentions identified
- The pattern matching system
  - 5,555 malignancy types extracted from NCI neoplasm ontology
  - Case-insensitive exact string matching applied
  - 85 malignancy type mentions **(42.1%)** recognized correctly
- The malignancy type extractor
  - 190 malignancy type mentions **(94.1%)** recognized correctly
  - Included all the baseline-identified mentions

Normalization

abdominal neoplasm
abdomen neoplasm
Abdominal tumour
Abdominal neoplasm NOS
Abdominal tumor
Abdominal Neoplasms
Abdominal Neoplasm
Neoplasm, Abdominal
Neoplasms, Abdominal
Neoplasm of abdomen
Tumour of abdomen
Tumor of abdomen
ABDOMEN TUMOR

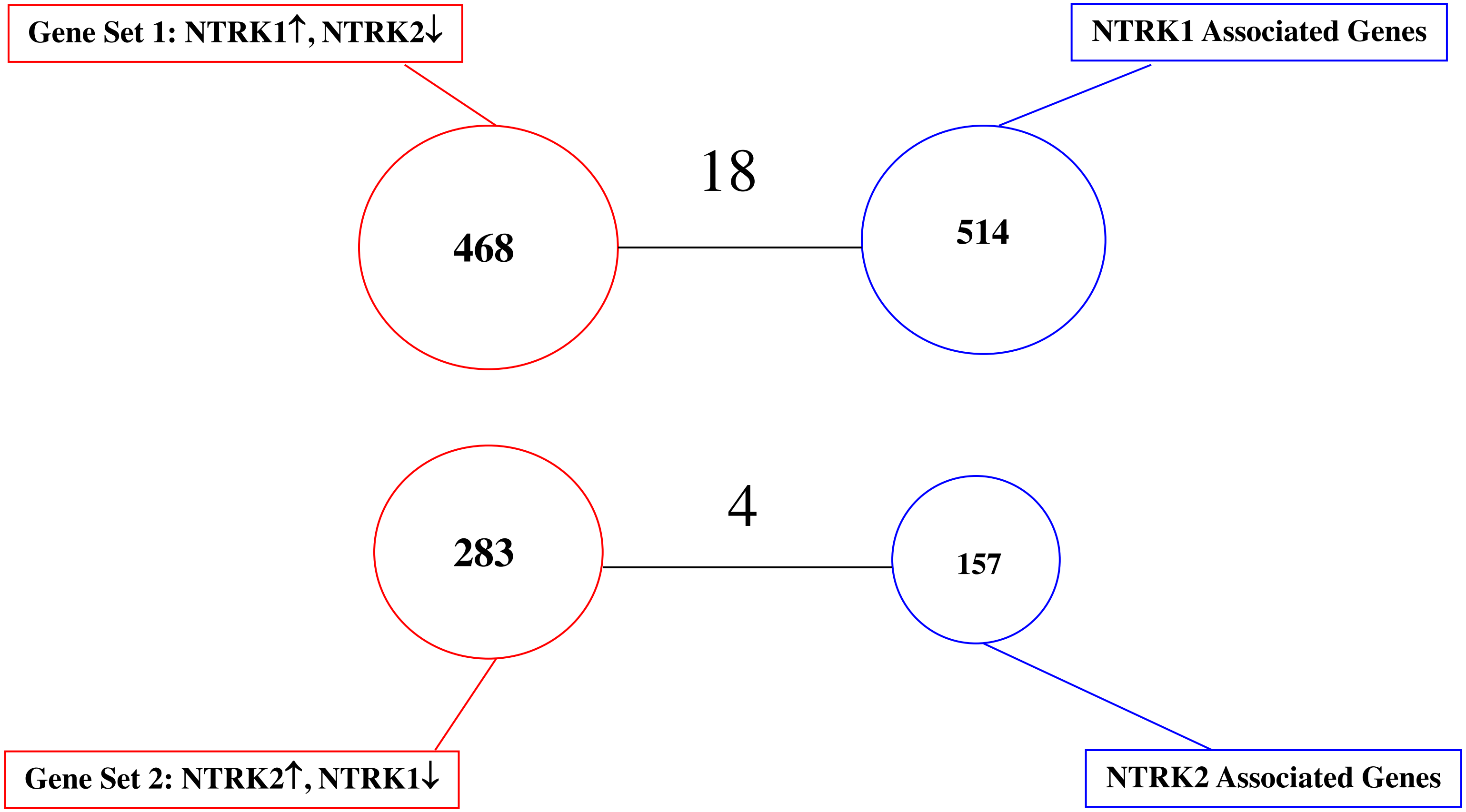
UMLS metathesaurus
Concept Unique Identifier (CUI)
19,397 CUIs with 92,414 synonyms

→ C0000735

Text Mining Applications -- Hypothesizing NB Candidate Genes

Microarray Expression Data Analysis

NTRK1/NTRK2 Associated Genes in Literature



Gene Set 1: NTRK1↑, NTRK2↓

NTRK1 Associated Genes

468

18

514

283

4

157

Gene Set 2: NTRK2↑, NTRK1↓

NTRK2 Associated Genes

Hypergeometric Test between Array and Overlap Groups

	Overlap Group
CD	<0.001
CGP	0.728
CCSI	0.00940
CM	0.0124
NSDF	<0.001
CAO	0.0117

Multiple-test corrected P-values (Bonferroni step-down)

Six selected pathways:

CD -- Cell Death;
CGP -- Cell Growth and Proliferation;
CCSI -- Cell-to-Cell Signaling and Interaction;
CM -- Cell Morphology;
NSDF -- Nervous System Development and Function;
CAO -- Cellular Assembly and Organization.

Ingenuity Pathway Analysis Tool Kit

Some personal history

- Prosody
  - Individuals are unsure, groups disagree
  - But … no "word constancy", maybe no phonology…
- Syntax
  - Individuals are unsure, groups disagree
  - But … categories and relations
    are part of theory of language itself
  - Thus, hard to separate "data" and "theory"
- Biomedical entities and relations
  - Individuals are unsure, groups disagree
  - … even though categories are external & consensual!
  - What's going on?

Perhaps this experience is telling us something about the nature of concepts and their extensions…

Why does this matter?

- The process is slow and expensive --
  ~6-18 months to converge
- The main roadblock is not the annotation itself,
  but the iterative development
  of annotation concepts and “case law”
- The results may be application-specific
  (or domain-specific)
- Despite conceptual similarities,
  generalization across applications
  has only been in human skill and experience,
  not in the core technology of statistical tagging

A blast from the past?

- This is like NL query systems ca. 1980, which worked well given ~1 engineer-year of adaptation to a new problem
- The legend: we've solved that problem
  - by using machine-learning methods
  - which don't need any new programming to be applied to a new problem
- The reality: it's just about as expensive
  - to manage the iterative development of annotation "case law"
  - and to create a big enough annotated training set
- Automated tagging technology works well
  - and many applications justify the cost
  - but the cost is still a major limiting factor

General solutions?

- Avoid human annotation entirely
  - Infer useful features from untagged text
  - Integrate other information sources
    (bioinformatic databases, microarray data, …)
- Pay the price -- once
  - Create a “basis set” of ready-made analyzers
    providing general solutions to the conceptual and linguistic issues
    … e.g. parser for biomedical text, ontology for biomedical concepts
  - Adapt easily to solve new problems

There are good ideas.

But so far, neither idea works well enough
to replace the iterative-refinement process
(rather than e.g. adding useful features
to supplement it)

A far-out idea

- An analogy to translation?
  - Entity/relation annotation is a (partial) translation from text into concepts
  - Some translations are really bad; some are better; but there is not one perfect translation -- instead we think of translation evaluation as some sort of distribution of a quality measure over an infinite space of word sequences
  - We don't try to solve MT by training translators to produce a unique output -- why do annotation that way?
- Perhaps we should evaluate (and apply) taggers in a way that accepts diversity rather than trying to eliminate it
- Umeda/Coker phrasing experiment…

Where are we?

- Goal is data

  … which we can use to develop/compare theories
- But “description is theory”

  … to some extent at least
- And even with shared theory

  (and language-external entities)

  achieving decent inter-annotator agreement
  requires a long process of “common law” development.

Suggestions

- Consider cost/benefit trade-offs
  - where *cost* includes
    - "common law" development time
    - annotator training time
    - and
  - and *benefit* includes
    - the resulting kappa
      (or other measure of information gain)
    - and the usefulness of the data
      for scientific exploration

FINIS

A farther-out idea

- Who is learning what?
  - A typical tagger is learning to map text features into b/i/o codes using a loglinear model.
  - A human, given the same series of texts with regions "highlighted", would try to find the simplest conceptual structure that fits the data (i.e. the simplest logical combination of primitive concepts)
  - The developers of annotation guidelines are simultaneously (and sequentially) choosing the text regions instantiating their current concept and revising or refining that concept
- If we had a good-enough proxy for the relevant human conceptual space (from an ontology, or from analysis of a billion words of text, or whatever), could we model this process?
  - what kind of "conceptual structures" would be learned?
  - via what sort of learning algorithm?
  - with what starting point and what ongoing guidance?