Multiply Annotated Corpora in Biomedical IE

Barry Haddow, Beatrice Alex
University of Edinburgh

LREC 2008, Marrakech
29 May 2008

Outline

1. Introduction
2. Dataset
3. System
4. Experiments
5. Conclusions

Why Multiple Annotation?

- Annotated data is used to train/test IE systems
- Normally multiply annotate a sample in order to
  - Verify that humans can reliably do the task
  - Provide a measure of difficulty
  - Monitor annotation quality
- Annotation is expensive
- Need to use the budget effectively
  - How much to multiply annotate?
  - How to use multiply annotated data?
  - Reconciliation expensive too — would like to avoid
- Annotator disagreement may represent real ambiguity.

Corpora

- The ITI TXM Corpora
- Two Corpora produced for TXM project
- Annotations of entities, enriched relations, normalisations
- Full papers from PubMed/PubMedCentral
- Split into TRAIN (64%), DEVTEST (16%) and TEST (20%)

PPI	TE
Protein-protein Interaction	Tissue Expression
75,000 sentences	60,000 sentences

Entity Annotations

Entity type	PPI	TE
CellLine	7,676	—
Complex	7,668	4,033
DevelopmentalStage	—	1,754
Disease	—	2,432
DrugCompound	11,886	16,131
ExperimentalMethod	15,311	9,803
Fragment	13,412	4,466
Fusion	4,344	1,459
GOMOP	—	4,647
Gene	—	12,059
Modification	6,706	—
mRNAcDNA	—	8,446
Mutant	4,829	1,607
Protein	88,607	60,782
Tissue	—	36,029

Relation Annotations

Corpus	Relation type	Count
PPI	PPI	11,523
PPI	FRAG	16,002
TE	TE	12,426
TE	FRAG	4,735

- PPI — Protein-protein interaction relations
- TE — Tissue expression relations
- FRAG — Fragment/Mutant - parent protein relations

Multiply Annotated Documents

Annotations	PPI	TE
Single	125	150
Double	65	86
Triple	27	2
Total documents	217	238
Total annotations	336	328

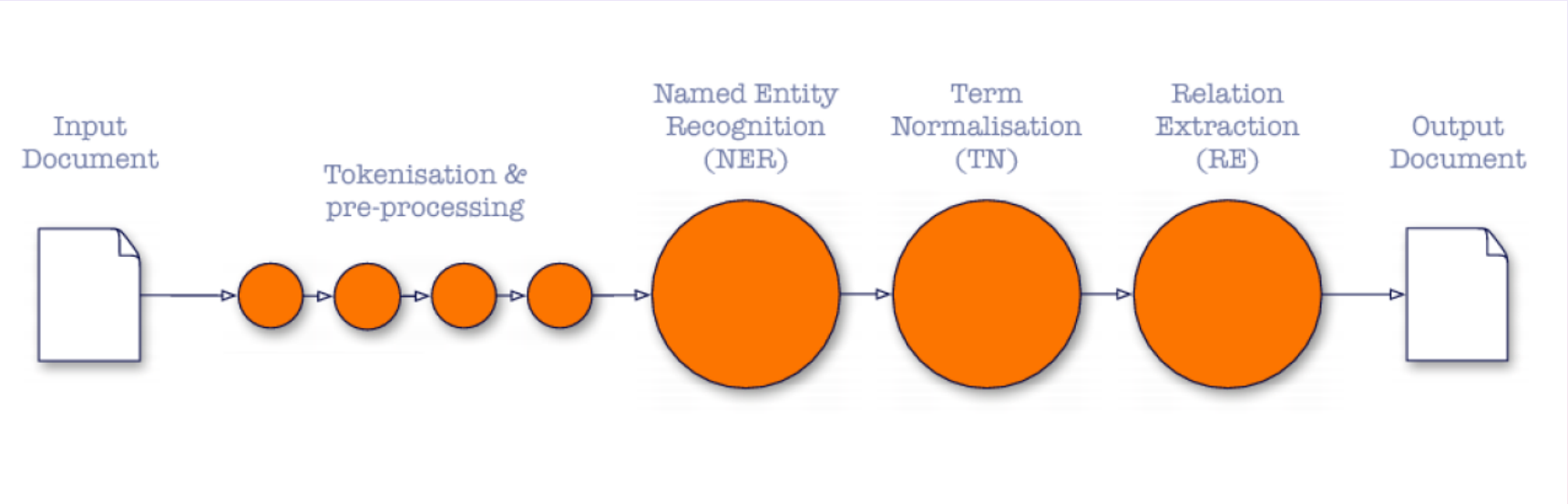
- A sample of documents from each corpus was multiply-annotated
- Most were doubly annotated, some triply

Inter-Annotator Agreement

Corpus	Annotated item	IAA
PPI	entities	84.9
PPI	relations	76.1
PPI	combined	59.7
TE	entities	83.8
TE	relations	74.1
TE	combined	55.7

- IAA measured using $F1$
- Micro-averaged across document pairs
- Combined is overall IAA for relations

TXM Pipeline



- Will focus on NER and RE here

Grover et al. 2007. Adapting a relation extraction pipeline for the BioCreAtIvE II task. BioCreAtIvE II Workshop 2007.

TXM

Named-entity Recognition (NER)

- Curran & Clark tagger: Maximum Entropy Markov Model tagger
- Extra features specifically for biomedical texts.
- Applied several biomedical gazetteers.
- Post-processing to correct boundary errors
- Detects nested entities
- Feature set optimised on DEVTEST
- *Alex at al. Recognising Nested Named Entities in Biomedical Text. BioNLP 2007.*

Relation Extraction (RE)

- Treats each relation type separately.
- Generates candidate relations as pairs of (gold) entities.
- Maximum entropy classifier chooses "yes" or "no"
- Features derived from entities, context, pos-tags, chunks etc.
- Feature set also optimised on DEVTEST
- *Haddow. Using Automated Feature Optimisation to Create an Adaptable Relation Extraction System. BioNLP 2008.*

Expt 1: How to use Multiply Annotated?

- How can we best use the multiply annotations in training?
- Try different methods for obtaining one annotated document(s) from set of equivalent documents.
- Choose Training Documents:
  - **all** — Use all documents
  - **one-random** — Pick one at random
  - **best-ner** — Highest ner score
  - **best-re** — Highest re score
  - **consistent** — Fixed annotator preference
- Combine Training Documents:
  - **intersection** — Entities/relations in both
  - **union** — Entities/relations in either
- Train with each strategy — test on DEVTEST and TEST

Results - NER

Method	PPI		TE	
	DEVTEST	TEST	DEVTEST	TEST
all	**75.1**	**72.5**	**65.4**	63.3
one-random	75.0	71.9	65.2	**63.7**
best-ner	74.7	72.1	64.9	63.5
best-re	**75.1**	72.1	65.1	63.6
consistent	74.7	72.4	65.0	63.6
intersection	74.8	72.0	63.9	62.8
union	74.8	72.1	65.1	**63.7**

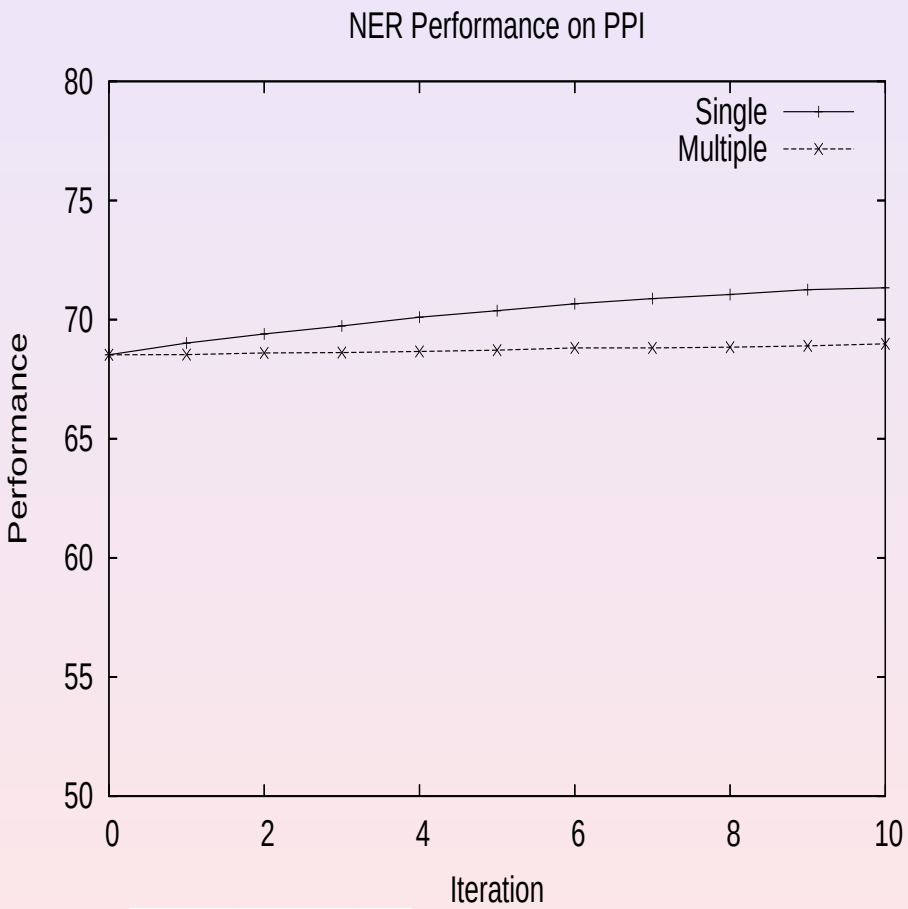
Results - RE

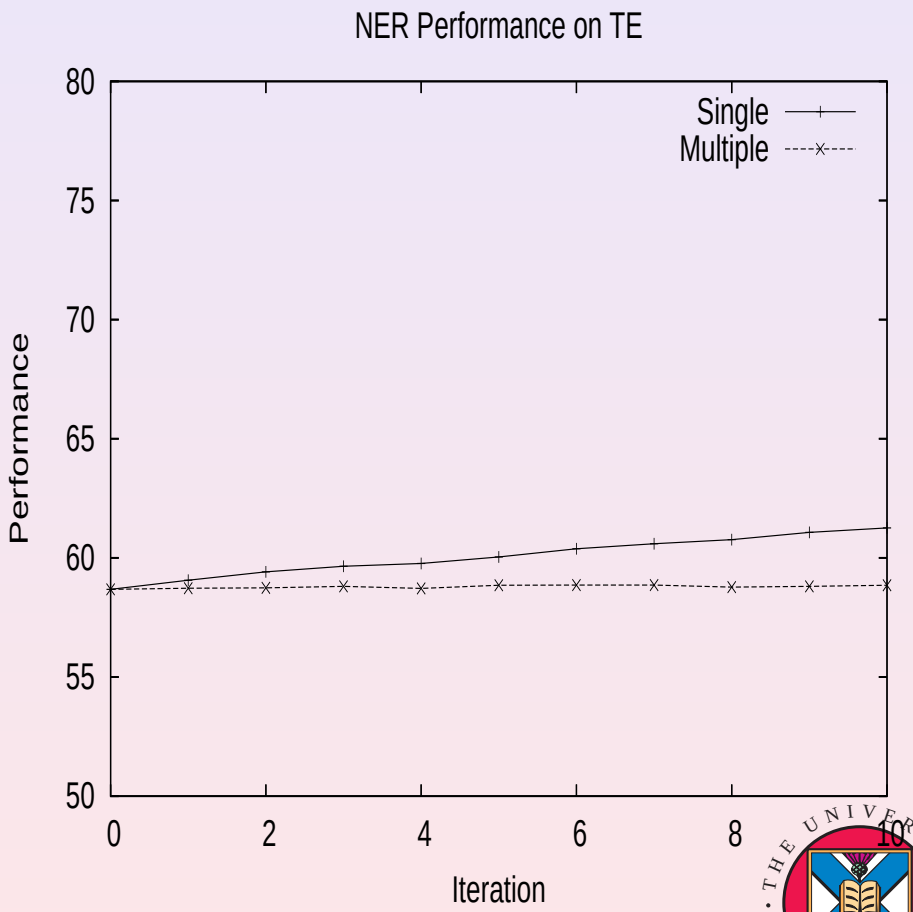
Method	PPI		TE	
	DEVTEST	TEST	DEVTEST	TEST
all	**58.9**	**58.7**	**58.3**	53.3
one-random	57.5	58.6	57.2	53.6
best-ner	57.9	58.5	57.0	**53.9**
best-re	57.7	58.6	57.0	52.8
consistent	57.8	58.6	57.0	53.4
intersection	56.9	58.3	54.6	53.1
union	56.4	58.0	56.4	53.1

Expt 2: Adding More Data

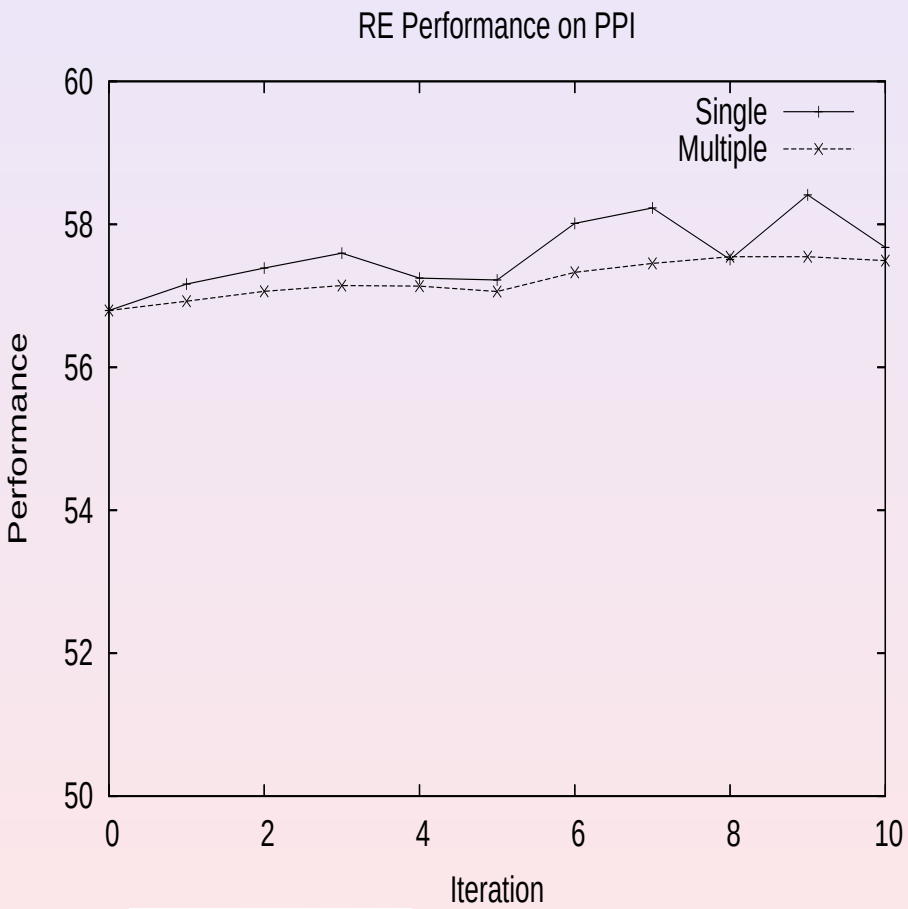
- Which gives the most rapid improvement in performance?
  - Add more singly annotated (new documents)
  - Add more multiply annotated
- Created learning curves for NER/RE, PPI/TE
- Started with fixed set of singly-annotated
- Added further documents in batches
  - Train/test in **all** configuration after each batch
- Randomise, repeat and average

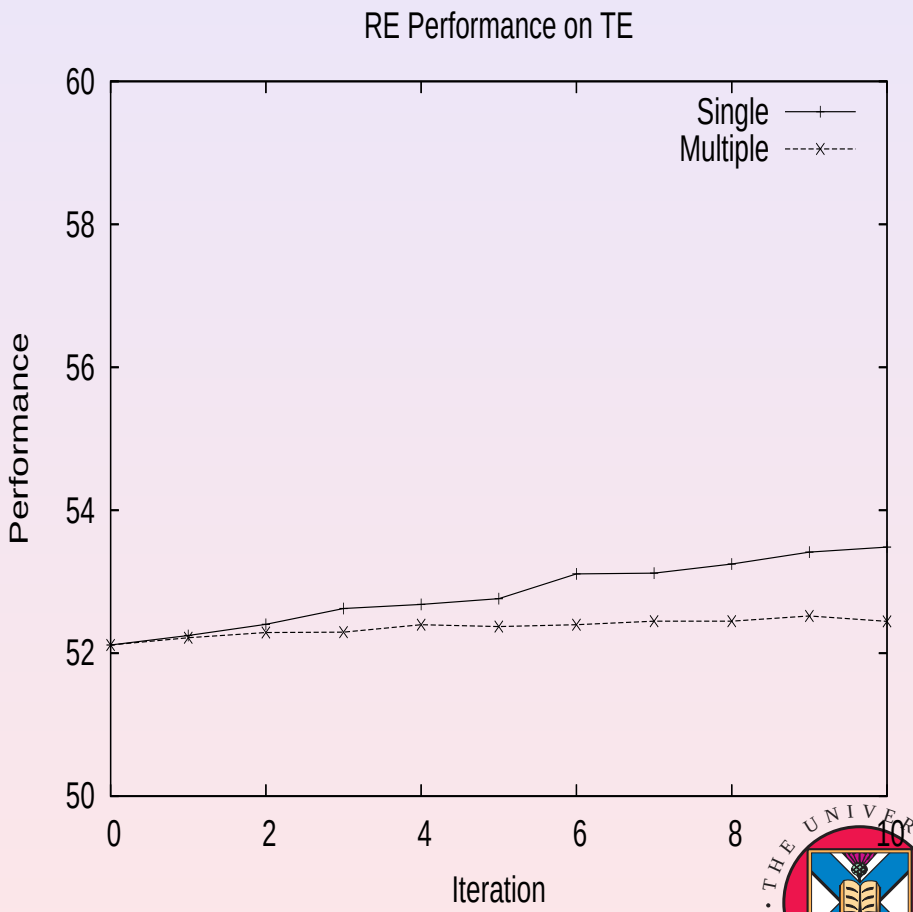
Results - NER


NER Performance on PPI
Performance
Iteration
Single
Multiple
80
75
70
65
60
55
50
0
2
4
6
8
10


NER Performance on TE
Performance
Iteration
Single
Multiple
80
75
70
65
60
55
50
0
2
4
6
8
10

Results - RE


RE Performance on PPI
Performance
Iteration
Single
Multiple
60
58
56
54
52
50
0
2
4
6
8
10


RE Performance on TE
Performance
Iteration
Single
Multiple
60
58
56
54
52
50
0
2
4
6
8
10

Experiment 1

- Using all annotated versions tends to give better results.
- However, no significant difference in combination methods
  - No gain from adding both annotated versions, as opposed to picking one.
  - For RE, an IAA below 60 means extra information in different annotated versions
  - But - large training corpus - inconsistency.
  - **union** and **intersection** strategies change precision-recall balance

Experiment 2

- Learning curves suggest extra singly-annotated is more useful
- Just annotate enough to give good estimate of IAA
- Lots of multiple annotation probably not useful for training

Acknowledgments

- ITI Life Sciences
- The Annotation Team
- The TXM Team: Claire Grover, Ewan Klein, Mijail Kabadjov, Michael Matthews, Stuart Roebuck, Richard Tobin and Xinglong Wang

Questions?

Thank you!
Questions?