

Unlocking astronomy knowledge from the literature

Tara Murphy and James R. Curran
School of IT and School of Physics
University of Sydney
Australia

20th February, 2008



Outline

- 1 Introduction
- 2 Natural Language Processing (NLP)
- 3 Astro Named Entity Recognition (Astro-NE)
- 4 Astro Bootstrapping (Astro-Boot)





We are undergoing an information explosion

- New surveys are generating tera-(and soon peta)-bytes of data
- Our processed datasets are extremely large and also rich and multidimensional



- More scientific publications being generated than ever before

A Medline search for *Breast Cancer treatment* returned 20 000 references in 2001. Now it returns 90 000

- **In some fields it is no longer possible for scientists to keep up with the literature**





Astronomy has excellent curated resources

- Astronomy is at the cutting edge of eScience
- We have *excellent* database, search and literature management tools
- For example
 - ADS
 - SIMBAD
 - NED
 - VO tools — e.g. Aladin
- We already have a very systematic approach to object naming and classification
- We also have the “sky” as a fundamental reference point





However there is still information we aren't exploiting

- Consider the following queries
 - Get all images of MySource?
 - What observations have been made of MySource?
 - What is the radio morphology of MySource?
 - What telescopes has MySource been observed with?
 - Does MySource have any properties of particular interest?
- This *semi-structured* data in the literature is not accessible automatically
- Semantic markup and ontologies are part of the solution



However there is still information we aren't exploiting

- Consider the following queries
 - Get all images of MySource?
 - What observations have been made of MySource?
 - What is the radio morphology of MySource?
 - What telescopes has MySource been observed with?
 - Does MySource have any properties of particular interest?
- This *semi-structured* data in the literature is not accessible automatically
- Semantic markup and ontologies are part of the solution

There will always be information that is only accessible **directly from the text** of scientific publications.





Search engines don't consider the meaning of words

- Search engines treat language as a *bag of words*
- Hence they doesn't realise that

(1) John loves Mary

(2) Mary loves John

do not mean the same thing, but that

(3) John, my cousin twice removed, loves Mary

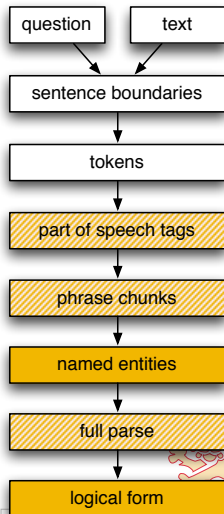
means the same as (1)

- Word proximity and order is necessary but still insufficient
- Search engines have no domain knowledge



The Natural Language Processing stack

- Each layer provides greater knowledge of the structure of the language
- Lower levels (striped)
 - syntactic structure
- Higher levels (solid)
 - semantic structure
- Each layer takes you further from the *bag of words* approach



Tagging, Chunking and Named Entities

Mr.	Vinken	is	chairman	of	Elsevier	N.V.
NNP	NNP	VBZ	NN	IN	NNP	NNP
I-NP	I-NP	I-VP	I-NP	I-PP	I-NP	I-NP
I-PER	I-PER	O	O	O	I-ORG	I-ORG

,	the	Dutch	publishing	group	.	tokens
,	DT	NNP	VBG	NN	.	POS tags
O	I-NP	I-NP	I-NP	I-NP	O	chunks
O	O	O	O	O	O	NE tags



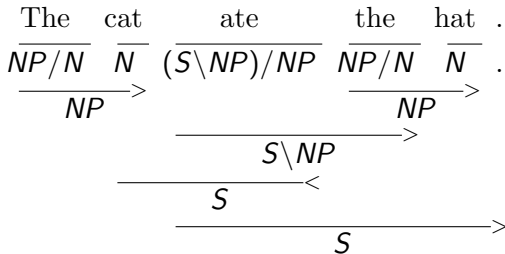


The performance of these tasks is high

- Part of Speech (POS) Tagging
 - 45 POS tags
 - Trained on 1 million words of the Penn Treebank text taken from the 1987 Wall Street Journal
 - 97% state of the art accuracy
- Named Entity (NE) Tagging
 - 9 NE tags
 - Trained on 160,000 words of Message Understanding Conference (MUC-7) data
 - 92 – 94% state of the art accuracy



Parsing is the process of identifying syntactic structure



- Parsing is slow
 - Parsing speed: 30 sentences/sec (500 words/sec)
 - NE tagging speed: 150 000 words/sec
- But even simple NLP is useful for information retrieval
- Clark and Curran 2007, *Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models*, Computational Linguistics, 33(4)





Scientific literature isn't the Wall Street Journal

- Almost all existing NLP tools are based on statistical models created on newspaper text
- We can't apply them directly to scientific text
- This is not surprising — scientific text has many special features, including a high level of jargon



Scientific literature isn't the Wall Street Journal

- Almost all existing NLP tools are based on statistical models created on newspaper text
- We can't apply them directly to scientific text
- This is not surprising — scientific text has many special features, including a high level of jargon

It was concluded that the carcinoembryonic antigens represent cellular constituents which are repressed during the course of differentiation of the normal digestive system epithelium and reappear in the corresponding malignant cells by a process of depressive dedifferentiation.





But they do have some things in common!

	Newsire	Astronomy
1	the	the
2	of	of
3	and	and
4	to	in
5	a	to
6	in	is
7	is	a
8	that	for
⋮	⋮	⋮

- The first specialized word comes in at No. 30



But they do have some things in common!

	Newsire	Astronomy
1	the	the
2	of	of
3	and	and
4	to	in
5	a	to
6	in	is
7	is	a
8	that	for
⋮	⋮	⋮

- The first specialized word comes in at No. 30

galaxies!





To use NLP tools we need an astronomy corpus

- The missing link is (lots of) training data
- We've developed the first large scale astronomy corpus
 - full-text extracted from \LaTeX papers in astro-ph (to 2005)
 - this is $\sim 52\,000$ articles or ~ 180 million words
 - equations transcribed approximately using Unicode
 - manually annotated with astronomy named entities
- Murphy, McIntosh and Curran 2006, *Named Entity Recognition for Astronomy Literature*, ALTW



We developed a set of Named Entity classes

Class	Definition	Examples	Comments
GXY	galaxy	NGC 4625; Milky Way; Galaxy	inc. black holes
GXYP	galaxy part	Local Bubble; 4C13.73	
GXYC	galaxy cluster	Abell 3266; Virgo Cluster	inc. small groups
NEB	nebula	Crab Nebula; Trapezium	
NEBP	nebula part	DR21(OH)	inc. star formation regions
STA	star	Mira A; PSR 0329+54; Sun	inc. pulsars
STAC	star cluster	M22; Palomar 13	
SUPA	supernova	SN1987A; SN1998bw	
PNT	planet	Earth; HD 11768 b; tau Boo	inc. extra-solar planets
MOO	moon	Moon; Io; S/2000 J2	
OBJ	object	IRS1; MMS1; Coma C	inc. objects of unknown type
EVT	cosmic event	GRB000911	inc. Gamma Ray Bursts
FRQ	frequency	10 Hz; 1.4 GHz	
WAV	wavelength	1.3 cm; 2800 Å	



Then set our annotators (PhD students) to work

```

emacs@mac-james.local
File Edit Options Buffers Tools QuASI Help
[Icons]
*LEGY and an angular resolution of 0.9°ANG .
The Tibet-IIIITEL array has operated since 1999IDAT with the same threshold and resolution [ 1 , 2 ] .
The ability of these arrays to detect gamma-ray sources has been demonstrated through detection of the
* CrabINEB NebulaINEB [ 3 ] and MrkIG 501IG [ 4 ] .
A wide angle survey conducted with the Tibet-IIIITEL HDITEL array found 19 directions with excesses gre
* ater than 4σ over the average background .
The TibetASy10 collaboration noted that these may be explained as statistical fluctuations , but one di
* rection corresponded to the CrabINEB NebulaINEB [ 1 ] .
We searched radio , optical , x-ray , and gamma-ray catalogues for corresponding objects within 1°ANG
* 6 of each of the TibetITEL directions , and we selected four promising candidate directions for observ
* ation with the WhippleITEL telescope during the 2001-2002IDAT season .
After receiving an update from the Tibet-IIIITEL all-sky survey , we selected one more candidate for ob
* servation during the 2002-2003IDAT season .
Table 1 gives a summary of the targets and observations .
The candidates were selected as follows : Tibet110BJ had a high significance and showed steady increas
* e through Tibet-IIIITEL HDITEL data .
Tibet910BJ is 0.3°ANG from a Seyfert 1 galaxy ( RGBIG J1337+243IG ) .
Tibet1410BJ is 0.7°ANG from an EGRETITEL unidentified ( 3EG10BJ J2021+371610BJ ) .
Tibet1610BJ had a high significance and is in the CygnusINEB starINEB fieldINEB .
Tibet055410BJ showed steady increase through Tibet-IIIITEL HDITEL and Tibet-IIIITEL and was second in s
* ignificance to the CrabINEB .
__TABLE__ TibetITEL gamma-ray source candidates .
__TABULAR__ CrabINEB included for reference .
Observations and Analysis .
Observations were made with the WhippleITEL 10 m gamma-ray telescope with the 490 pixel camera [ 5 ] .
Only the inner 379 pixels ( 2.4°ANG ) were used in the analysis , and because of the large uncertain
* ty in the TibetITEL ASITEL source coordinates , 2-dimensional ( 2-D ) analysis was required .
In the analysis , potential gamma rays were selected by applying standard supercuts shape cuts ,
Then the distance and alpha cuts ( α < 10° ) were applied across a grid of points , with alpha and di
* stance calculated with respect to each point .
-u:-- 0305585.ann 22% L20 SVN-37 (QuASI annotation)-----

```



An example of the annotated text

Our **FUSE**_{|TEL} spectrum of **HD**_{|STA} **73882**_{|STA} is derived from time-tagged observations over the course of 8 orbits on **1999**_{|DAT} **Oct**_{|DAT} **30**_{|DAT} . Several “burst” events occurred during the observation (**Sahnow**_{|PER} et al. **2000**_{|DAT}) . We excluded all **photon**_{|PART} events that occurred during the bursts , reducing effective on-target integration time from **16.8**_{|DUR} **ksec**_{|DUR} to **16.1**_{|DUR} **ksec**_{|DUR} . Strong interstellar extinction and lack of co-alignment of the SiC channels with the LiF channels prevented the collection of useful data shortward of **1010**_{|WAV} **AA**_{|WAV} .



Our initial results for NE recognition

- We have manually annotated a corpus of 200 000 words
- 43 entity types (more than most comparable corpora)
- Inter-annotator agreement on subset of 30 000 words (96%)
- 10-fold cross validation
- 88% F-score on fine-grained categories
- 91% F-score on coarse-grained categories
- F-score is harmonic mean of precision and recall:

$$F = \frac{2PR}{P + R} \quad (1)$$

- F-score balances both precision and recall





So what can we do with this?

- Improving literature search tools by filtering or preferring segments that contain sought entity types
- Extracting summary information (such as morphologies)
- Helping to (semi-)automatically populate ontologies/gazetteers
- We extended other NLP techniques for populating ontologies. . .



Extracting semantic categories with bootstrapping

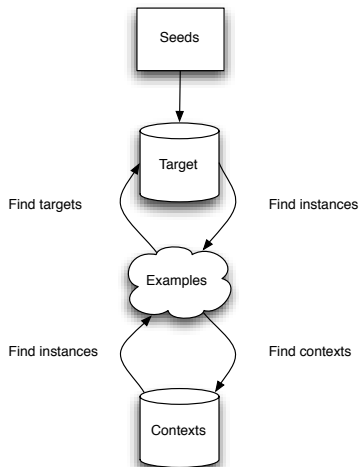
- We use indicative contexts to extract similar terms
- And similar terms to extract more indicative contexts
- For example, this context is indicative of telescopes:

observed	with	the	Hubble	telescope
X	X	X	term	X

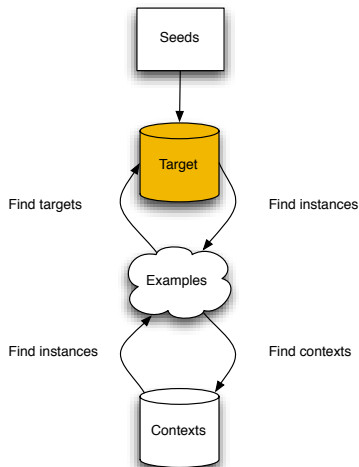
- Any context where **Hubble** occurs *may* be indicative of other telescopes
- Curran, Murphy and Scholz, 2007 *Minimising Semantic Drift Using Mutual Exclusion Bootstrapping*, PACLING
- Murphy and Curran, 2007 *Experiments in Mutual Exclusion Bootstrapping*, ALTW



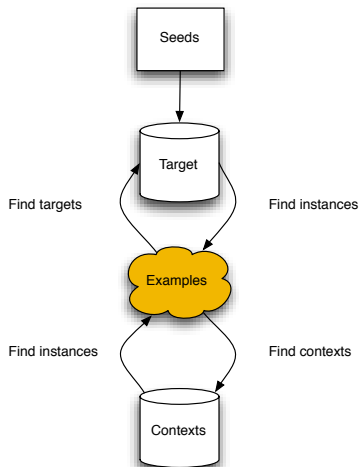
Mutual Bootstrapping (Riloff and Jones, 1999)



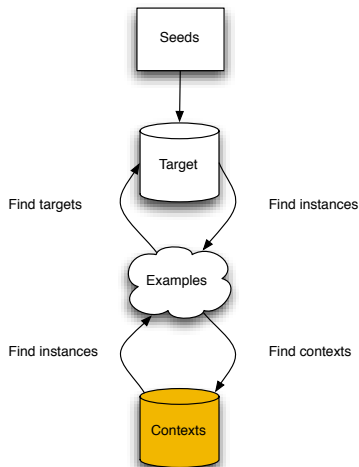
Mutual Bootstrapping (Riloff and Jones, 1999)



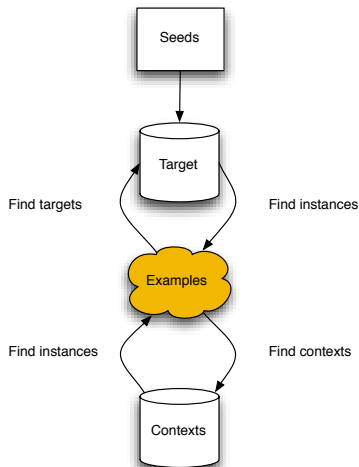
Mutual Bootstrapping (Riloff and Jones, 1999)



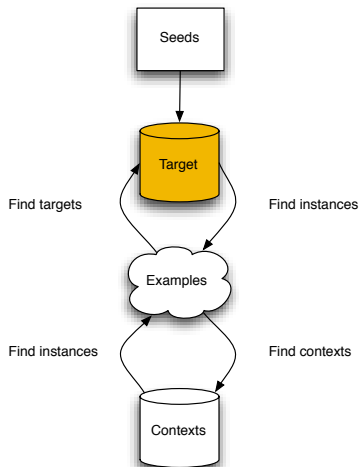
Mutual Bootstrapping (Riloff and Jones, 1999)



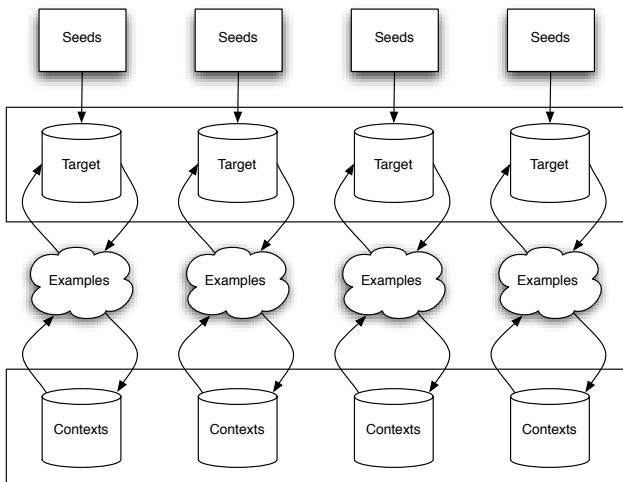
Mutual Bootstrapping (Riloff and Jones, 1999)



Mutual Bootstrapping (Riloff and Jones, 1999)



Mutual Exclusion Bootstrapping





Summary

- There is a lot of information “locked away” in scientific publications
- NLP techniques can help us extract this
- These could be incorporated into many aspects of the VO and existing astronomy services (e.g. ADS)
- NLP complements more structured approaches

- This project has been supported by the Australian Research Council under Discovery Project DP0665973.

