Named Entity Recognition for Astronomy Literature

Tara Murphy and **Tara McIntosh** and **James R. Curran**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{tm,tara,james}@it.usyd.edu.au

Abstract

We present a system for named entity recognition (NER) in astronomy journal articles. We have developed this system on a NE corpus comprising approximately 200,000 words of text from astronomy articles. These have been manually annotated with ∼40 entity types of interest to astronomers.

We report on the challenges involved in extracting the corpus, defining entity classes and annotating scientific text. We investigate which features of an existing state-of-the-art Maximum Entropy approach perform well on astronomy text. Our system achieves an F-score of 87.8%.

1 Introduction

Named entity recognition (NER) involves assigning broad semantic categories to entity references in text. While many of these categories do in fact refer to *named* entities, e.g. PERSON and LOCATION, others are not proper nouns, e.g. DATE and MONEY. However, they are all syntactically and/or semantically distinct and play a key role in Information Extraction (IE). NER is also a key component of Question Answering (QA) systems (Hirschman and Gaizauskas, 2001). State-of-the-art QA systems often have custom-built NER components with finer-grained categories than existing corpora (Harabagiu et al., 2000). For IE and QA systems, generalising entity references to broad semantic categories allows shallow extraction techniques to identify entities of interest and the relationships between them.

Another recent trend is to move beyond the traditional domain of newspaper text to other corpora. In particular, there is increasing interest in extracting information from scientific documents, such as journal articles, especially in biomedicine (Hirschman et al., 2002).

A key step in this process is understanding the entities of interest to scientists and building models to identify them in text. Unfortunately, existing models of language perform very badly on scientific text even for the categories which map directly between science and newswire, e.g. PERSON. Scientific entities often have more distinctive orthographic structure which is not exploited by existing models.

In this work we identify entities within astronomical journal articles. The astronomy domain has several advantages: firstly, it is representative of the physical sciences; secondly, the majority of papers are freely available in a format that is relatively easy to manipulate (LaTeX); thirdly, there are many interesting entity types to consider annotating; finally, there are many databases of astronomical objects that we will eventually exploit as gazetteer information.

After reviewing comparable named entity corpora, we discuss aspects of astronomy that make it challenging for NLP. We then describe the corpus collection and extraction process, define the named entity categories and present some examples of interesting cases of ambiguity that come up in astronomical text.

Finally, we describe experiments with retraining an existing Maximum Entropy tagger for astronomical named entities. Interestingly, some feature types that work well for newswire significantly degrade accuracy here. We also use the tagger to detect errors and inconsistencies in the annotated corpus. We plan to develop a much larger freely available astronomy NE corpus based on our experience described here.

2 Existing annotated corpora

Much of the development in NER has been driven by the corpora available for training and evaluating such systems. This is because the state-of-the-art systems rely on statistical machine learning approaches.

2.1 Message Understanding Conference

The MUC named entity recognition task (in MUC 6/7) covered three types of entities:

names PERSON, LOCATION, ORGANISATION;

temporal expressions DATE, TIME;

numeric expressions MONEY, PERCENT.

The distribution of these types in MUC 6 was: names 82%, temporal 10% and numeric 8%, and in MUC 7 was: names 67%, temporal 25% and numeric 6%.

The raw text for the MUC 6 NER corpus consisted of 30 Wall Street Journal articles, provided by the Linguistic Data Consortium (LDC). The text used for the English NER task in MUC 7 was from the New York Times News Service, also from the LDC. There are detailed annotation guidelines available.[1]

2.2 GENIA corpus

The GENIA corpus (Kim et al., 2003) is a collection of 2000 abstracts from the National Library of Medicine's MEDLINE database. The abstracts have been selected from search results for the keywords human, blood cells and transcription factors. GENIA is annotated with a combination of part of speech (POS) tags based on the Penn Treebank set (Marcus et al., 1994) and a set of biomedical named entities described in the GENIA ontology. One interesting aspect of the GENIA corpus is that some named entities are syntactically nested. However, most statistical NER systems are sequence taggers which cannot easily represent hierarchical tagging.

2.3 Astronomy Bootstrapping Corpus

The Astronomy Bootstrapping Corpus (Becker et al., 2005; Hachey et al., 2005) is a small corpus consisting of 209 abstracts from the NASA Astronomical Data System Archive. The corpus was developed as part

---

[1] www.cs.nyu.edu/cs/faculty/grishman/muc6.html

of experiments into efficient methods for developing new statistical models for NER. The abstracts were selected using the query quasar + line from articles published between 1997 and 2003. The corpus was annotated with the following named entity types:

1. INSTRUMENT NAME (136 instances)
2. SOURCE NAME (111 instances)
3. SOURCE TYPE (499 instances)
4. SPECTRAL FEATURE (321 instances)

The seed and test sets (50 and 159 abstracts) were annotated by two astronomy PhD students. The abstracts contained on average 10 sentences with an average length of 30 tokens, implying an *tag density* (the percentage of words tagged as a named entity) of $\sim 2\%$.

3 NLP for astronomy

Astronomy is a broad scientific domain combining theoretical, observational and computational research, which all differ in conventions and jargon. We are interested in NER for astronomy within a larger project to improve information access for scientists.

There are several comprehensive text and scientific databases for astronomy. For example, NASA Astrophysics Data System (ADS, 2005) is a bibliographic database containing over 4 million records (journal articles, books, etc) covering the areas of astronomy and astrophysics, instrumentation, physics and geophysics. ADS links to various external resources such as electronic articles, data catalogues and archives.

3.1 IAU naming conventions

The naming of astronomical objects is specified by the International Astronomical Union's (IAU) Commission 5, so as to minimise confusing or overlapping designations in the astronomical literature. The most common format for object names is a catalogue code followed by an abbreviated position (Lortet et al., 1994). Many objects still have common or historical names (e.g. the Crab Nebula). An object that occurs in multiple catalogues will have a separate name in each catalogue (e.g. PKS 0531+21 and NGC 1952 for the Crab Nebula).

1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
2	90	218	421	1909	3221	4320	5097	5869	6361	6556	7367	7732	3495

Table 1: Number of LaTeX astro-ph articles extracted for each year.

3.2 The Virtual Observatory

There is a major effort in astronomy to move towards integrated databases, software and telescopes. The umbrella organisation for this is the *International Virtual Observatory Alliance* (Hanisch and Quinn, 2005). One of the aims is to develop a complete ontology for astronomical data which will be used for Unified Content Descriptors (Martinez et al., 2005).

4 Collecting the raw corpus

The process of collecting the raw text for named entity annotation involved first obtaining astronomy text, extracting the raw text from the document formatting and splitting it into sentences and tokens.

4.1 arXiv

arXiv (arXiv, 2005) is an automated distribution system for research articles, started in 1991 at the Los Alamos National Laboratory to provide physicists with access to prepublication materials. It rapidly expanded to incorporate many domains of physics and thousands of users internationally (Ginsparg, 2001).

The astrophysics section (astro-ph) is used by most astronomers to distribute papers before or after publication in a recognised journal. It contains most of astrophysics publications from the last five to ten years. Table 1 shows that the number of articles submitted to astro-ph has increased rapidly since 1995.

The articles are mostly typeset in LaTeX. We have downloaded 52 658 articles from astro-ph, totalling approximately 180 million words. In creating the NE corpus we limited ourselves to articles published since 2000, as earlier years had irregular LaTeX usage.

4.2 LaTeX conversion

After collecting of LaTeX documents, the next step was to extract the text so that it could be processed using standard NLP tools. This normally involves ignoring most formatting and special characters in the documents.

However formatting and special characters play a major role in scientific documents, in the form of mathematical expressions, which are interspersed through the text. It is impossible to ignore every non-alphanumeric character or or map them back to some standard token because too much information is lost. Existing tools such as DeTeX (Trinkle, 2002) remove LaTeX markup including the mathematics, rendering scientific text nonsensical. Keeping the LaTeX markup is also problematic since the tagger's morphological features are confused by the markup.

4.3 LaTeX to Unicode

Our solution was to render as much of the LaTeX as possible in text, using Unicode (Unicode Consortium, 2005) to represent the mathematics as faithfully as possible. Unicode has excellent support for mathematical symbols and characters, including the Greek letters, operators and various accents.

Mapping LaTeX back to the corresponding Unicode character is difficult. For example, `\acirc`, `\^{a}` and `\hat{a}` are all used to produce â, which in Unicode is 0x0174.

Several systems attempt to convert LaTeX to other formats e.g. XML (Grimm, 2003). No existing system rendered the mathematics faithfully enough or with high enough coverage for our purposes. Currently our coverage of mathematics is very good but there are still some expressions that cannot be translated, e.g. complex nested expressions, rare symbols and non-Latin/Greek/Hebrew alphabetic characters.

4.4 Sentences and Tokenisation

We used MXTerminator (Reynar and Ratnaparkhi, 1997) as the sentence boundary detector with an additional Python script to fix common errors, e.g. mistaken boundaries on Sect. and et al. We used the Penn Treebank (Marcus et al., 1994) sed script to tokenize the text, again with a Python script to fix common errors, e.g. splitting numbers like 1,000 on the

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 **Å**|WAV .

Figure 1: An extract from our final corpus, originally from astro-ph/0005090.

comma, and reattaching the L^AT_EX which the tokenizer split off incorrectly.

4.5 The Corpus

The articles for the corpus were selected randomly from the downloaded LaTeX documents and annotated by the second author. The annotation was performed using a custom Emacs mode which provided syntax highlighting for named entity types and mapped the keys to specific named entities to make the annotation as fast as possible. The average annotation speed was 165 tokens per minute. An extract from our corpus is shown in Figure 1. There are a total of 7840 sentences in our corpus, with an average of 26.1 tokens per sentence.

5 Named Entity Categories

Examples of the categories we used are listed in Table 2. We restricted ourselves to high level categories such as STAR and GALAXY rather than detailed ones typically used by astronomers to classify objects such as red giant and elliptical galaxy.

5.1 Areas of Ambiguity

There were some entities that did not clearly fit into one specific category or are used in a way that is ambiguous. This section outlines some of these cases.

Temperature and Energy Due to the high temperatures in X-ray astronomy, temperatures are conventionally referred to in units of energy (eV), for example:

its 1 MeV temperature, the emission from...

Our annotation stays *consistent to the units*, so these cases are tagged as energies (EGY).

Angular distance Astronomers commonly refer to angular distances on the sky (in units of arc) because it is not possible to know the true distance between two objects without knowing their redshift. We annotate these according to the units, i.e. angles, although they are often used in place of distances.

Spectral lines and ions Absorption or emission of radiation by ions results in *spectral line features* in measured spectra. Common transitions have specific names (e.g. Hα) whereas others are referred to by the ion name (e.g. Si IV), introducing ambiguity.

5.2 Comparison with GENIA and MUC

The corpus has a named entity density of 5.4% of tokens. This is significantly higher than the density of the Astronomy Bootstrapping Corpus. The most frequency named entities types are: PER (1477 tags), DAT (1053 tags), TEL (867 tags), GAL (551 tags), and WAV (451 tags). The token 10 has the highest degree of ambiguity since it was tagged with every unit related tag: EGY, MASS, etc. and also as OBJ.

By comparison the GENIA corpus has a much higher density of 33.8% tokens on a sample the same size as our corpus. The highest frequency named entity types are: OTHER (16171 tags), PROTEIN (13197 tags), DNA DOMAIN (6848 tags) and PROTEIN FAMILY (6711 tags).

The density of tags in MUC is 11.8%, higher than our corpus but much lower than GENIA. The highest frequency named entities are OR-GANISATION (6373 tags), followed by LOCA-TION (3828 tags) and DATE (3672 tags). Table 3 gives a statistical comparison of the three corpora. This data suggests that the astronomy data will be harder to automatically tag than MUC 7 because the density is lower and there are many more classes. However, if there were more classes or finer grained distinctions in MUC this would not necessarily be true. It also demonstrates how different biological text is to other scientific domains.

Class	Definition	Examples	Comments
GXY	galaxy	NGC 4625; Milky Way; Galaxy	inc. black holes
NEB	nebula	Crab Nebula; Trapezium	
STA	star	Mira A; PSR 0329+54; Sun	inc. pulsars
STAC	star cluster	M22; Palomar 13	
SUPA	supernova	SN1987A; SN1998bw	
PNT	planet	Earth; Mars ; HD 11768 b; tau Boo	inc. extra-solar planets
FRQ	frequency	10 Hz; 1.4 GHz	
DUR	duration	13 seconds; a few years	inc. ages
LUM	luminosity	10^{46} ergs^{-1}; 10^{10}L$_\odot$	inc. flux
POS	position	17:45.6; -18:35:31; $17^h12^m13^s$	
TEL	telescope	ATCA; Chandra X-ray observatory	inc. satellites
ION	ion	Si IV; HCO$^+$	inc. molecular ions
SUR	survey	SUMSS; 2 Micron All-Sky Survey	
DAT	date	2003; June 17; 31st of August	inc. epochs (e.g. 2002.7)

Table 2: Example entity categories.

Corpus	ASTRO	GENIA	MUC
# cats	43	36	8
# entities	10 744	40 548	11 568
# tagged	16,016	69 057	19 056
# avg len	1.49	1.70	1.64
tag density	5.4%	33.8%	11.8%

Table 3: Comparison with GENIA and MUC.

Condition	Contextual predicate
$f(w_i) < 5$	X is prefix of w_i, $|X| \leq 4$
	X is suffix of w_i, $|X| \leq 4$
	w_i contains a digit
	w_i contains uppercase character
	w_i contains a hyphen
$\forall w_i$	$w_i = X$
	$w_{i-1} = X$, $w_{i-2} = X$
	$w_{i+1} = X$, $w_{i+2} = X$
$\forall w_i$	$\text{POS}_i = X$
	$\text{POS}_{i-1} = X$, $\text{POS}_{i-2} = X$
	$\text{POS}_{i+1} = X$, $\text{POS}_{i+2} = X$
$\forall w_i$	$\text{NE}_{i-1} = X$
	$\text{NE}_{i-2}\text{NE}_{i-1} = XY$

Table 4: Baseline contextual predicates

6 Maximum Entropy Tagger

The purpose of creating this annotated corpus is to develop a named entity tagger for astronomy literature. In these experiments we adapt the C&C NE tagger (Curran and Clark, 2003) to astronomy literature by investigating which feature types improve the performance of the tagger. However, as we shall see below, the tagger can also be used to test and improve the quality of the annotation. It can also be used to speed up the annotation process by pre-annotating sentences with their most likely tag. We were also interested to see whether > 40 named-entity categories could be distinguished successfully with this quantity of data.

Condition	Contextual predicate
$f(w_i) < 5$	w_i contains period/punctuation
	w_i is only digits
	w_i is a number
	w_i is {upper,lower,title,mixed} case
	w_i is alphanumeric
	length of w_i
	w_i has only Roman numerals
	w_i is an initial (X.)
	w_i is an acronym (ABC, A.B.C.)
$\forall w_i$	memory NE tag for w_i
	unigram tag of w_{i+1}, w_{i+2}
$\forall w_i$	w_i, w_{i-1} or w_{i+1} in a gazetteer
$\forall w_i$	w_i not lowercase and $f_{lc} > f_{uc}$
$\forall w_i$	uni-, bi- and tri-grams of word type

Table 5: Contextual predicates in final system

The C&C NE tagger feature types are shown in Tables 4 and 5. The feature types in Table 4 are the same as used in MXPost (Ratnaparkhi, 1996) with the addition of the NE tag history features. We call this the *baseline* system. Note, this is not the baseline of the NE tagging task, only the baseline performance for a Maximum Entropy approach.

Table 5 includes extra feature types that were tested by Curran and Clark (2003). The w_i *is only digits* predicates apply to words consisting of all digits. Title-case applies to words with an initial uppercase letter followed by lowercase (e.g. `Mr`). Mixed-case applies to words with mixed lower- and uppercase (e.g. `CitiBank`). The length features encode the length of the word from 1 to 15 characters, with a single bin for lengths greater than 15.

The next set of contextual predicates encode extra information about NE tags in the current context. The memory NE tag predicate records the NE tag that was most recently assigned to the current word. This memory is reset at

N	Word	Correct	Tagged
23	OH	MOL	NONE
14	rays	PART	NONE
8	GC	GXYP	NONE
6	cosmic	PART	NONE
6	HII	ION	NONE
5	telescope	TEL	NONE
5	cluster	STAC	NONE
5	and	LUM	NONE
4	gamma	NONE	PART

Table 6: Detected Errors and Ambiguities

Experiment	P	R	F-score
BASELINE	93.0	82.5	87.5
EXTENDED	91.2	82.4	86.6
-MEMORY	92.1	84.3	88.0
-MEMORY/POS	92.3	83.9	87.9
COARSE BASE	92.6	86.7	89.5
COARSE EXTENDED	93.0	88.9	90.9

Table 7: Feature experiment results

the beginning of each document. The unigram predicates encode the most probable tag for the next words in the window. The unigram probabilities are relative frequencies obtained from the training data. This feature enables us to know something about the likely NE tag of the next word before reaching it.

Another feature type encodes whether the current word is more frequently seen in lowercase than title-case in a large external corpus. This is useful for disambiguating beginning of sentence capitalisation. Eventually the frequency information will come from the raw astronomy corpus itself.

Collins (2002) describes a mapping from words to *word types* which groups words with similar orthographic forms into classes. This involves mapping characters to classes and merging adjacent characters of the same type. For example, Moody becomes Aa, A.B.C. becomes A.A.A. and 1,345.05 becomes 0,0.0. The classes are used to define unigram, bigram and trigram contextual predicates over the window. This is expected to be a very useful feature for scientific entities.

7 Detecting Errors and Ambiguities

We first trained the C&C tagger on the annotated corpus and then used this model to retag the corpus. We then compared this retagged corpus with the original annotations. The differences were manually checked and corrections made where necessary.

Table 6 shows the most frequent errors and ambiguities detected by this approach. Most of the differences found were either the result of genuine ambiguity or erroneous annotation.

GC, cosmic and HII are examples of genuine ambiguity that is difficult for the tagger to model correctly. GC means globular cluster which

is not tagged, but less often refers to the Galactic Centre which *is* tagged (GXYP). cosmic occurs in two contexts: as part of cosmic ray(s) which is tagged as a particle; and in expressions such as cosmic microwave background radiation which is not tagged. HII is used most frequently in reference to HII ions and hence is tagged as an (ION). However, occasionally HII is used to refer to HII galaxies and not tagged.

OH and gamma rays are examples where there was some inconsistency or error in some of the annotated data. In both of these cases instances in the corpus were not tagged.

We also implemented the approach of Dickinson and Meurers (2003) for identifying annotation errors in part of speech (POS) tagging. Their approach finds the longest sequence of words that surround a tagging ambiguity. The longer the context, the more likely the ambiguity is in fact an annotation error. This approach identified a number of additional errors, particularly annotation errors within entities. However, many of the errors we may have found using this technique were already identified using the tagging described above.

8 Inter-annotator Agreement

To test the reliability of the annotations we performed two tests. Firstly, we asked an astronomy PhD student to take our annotation guidelines and annotate around 30,000 words (15% of the corpus). Secondly, the second author also reannotated a different 30,000 words about 2-months after the original annotation process to check for self consistency.

We used the kappa statistic (Cohen, 1960) to evaluate inter-annotator reliability. The kappa value for agreement with the PhD student annotation was 0.91 on all tags and 0.82 not including the NONE tags. Given that the annotation guidelines were not as complete as we would have liked, this agreement is very

good. The kappa value for agreement with the reannotated corpus was 0.96 on all tags and 0.92 not including the NONE tags.

When the differences between the 30,000 word sections and the original corpus were check manually (by the second author and the PhD student) practically all of them were found to be annotation errors rather than genuine ambiguity that they could not agree on.

9 Results

We split the corpus into 90% training and 10% testing sets. For our final results we performed 10-fold cross validation. For the experiments analysing the contribution of named entity feature types from the C&C tagger we used one of the 10 folds. The evaluation was performed using the CoNLL shared task evaluation script[1].

9.1 Feature Experiments

The results of the feature experiments are shown in Table 7. The Maximum Entropy baseline performance of 87.5% F-score is very high given the large number of categories. Clearly there is enough contextual information surrounding the entities that they can be fairly reliably tagged.

A surprising result is that using all of the additional features which helped significantly improve performance on newswire actually damages performance by $\sim 1\%$. Further experimental analysis with removing specific feature types found that the offending feature was the last tagged with tag x feature (the *memory* feature). Removing this feature improves performance a little bit more giving our best result of 88.0% F-score. We believe this feature performs particularly badly on numeric expressions which are part of many different named entity classes which may appear with the same word in a single article.

We experimented with removing the POS tag features since the POS tagger performed very badly on astronomy text, but this made little difference. We have experimented with removing the other feature types listed in Table 5 but this resulted in a small decrease in performance each time.

This demonstrates that with new training data it is fairly straightforward to achieve rea-

Category	Constituent categories
galaxy	GXY, GXYP, GXYC, NEBP, NEB
star	STA, STAP, STAC, SUPA
object	OBJ, OBJP, EVT
sso	PNT, PNTP, MOO, MOOP
units	FRQ, WAV, DIST, TEMP, DUR, MASS, ANG, LUM, VEL, PCT, EGY, UNIT, POS
inst.	TEL, INST
particle	PART, ELEM, MOL, ION, LN
person	PER, ORG, URL
location	LOC
obs.	SUR, CAT, DB
date	DAT, TIME
software	CODE

Table 8: Coarse-grained mapping

sonable performance in identifying astronomy named entities.

9.2 Coarse-grained categories

One interesting property of our named entity corpus is the very large number of categories relative to existing NE corpora such as MUC. To test what impact the number of classes has on performance we repeated the experiment described above, using coarser-grained named entity categories based on the mapping shown in Table 8.

The course grained classifier achieves an F-score of 89.5% using the baseline feature set and an F-score of 90.9% using the extended feature set without the memory feature. The key difference between the fine and coarse grained results is the significantly better recall on coarse grained classes.

10 Conclusion

This is a pilot annotation of astronomy texts with named entity information. Now that we have created the initial corpus we intend to reevaluate the categories, aiming for greater consistency and coverage of the entities of interest in the corpus.

We have performed preliminary experiments in training taggers using our corpus. These experiments have produced very promising results so far (87.8% F-score on 10-fold cross validation). We intend to extend our evaluation of individual features for scientific text and add features that exploit online astronomy resources.

This paper has described in detail the process of creating a named entity annotated corpus of astronomical journal articles and conference papers. This includes translating the

LATEX typesetting information into a useable format. Unlike existing work we have rendered the mathematics in Unicode text rather than just removing it, which is important for further analysis of the data. The resulting corpus is larger than existing resources, such as MUC, but has been annotated with a much more detailed set of over 40 named entity classes.

Finally, we have demonstrated that high accuracy named entity recognisers can be trained using the initial release of this corpus, and shown how the tagger can be used to iteratively identify potential tagging errors. The quality of the results should only improve as the corpus size and quality is increased.

11   Acknowledgments

References

ADS. 2005. Astronomical Data Service. `http://www.adsabs.harvard.edu/`.

arXiv. 2005. arXiv.org archive. `http://arxiv.org`.

M. Becker, B. Hachey, B. Alex, and C. Grover. 2005. Optimising selective sampling for bootstrapping named entity recognition. In *Proceedings of the ICML Workshop on Learning with Multiple Views*, pages 5–11, Bonn, Germany.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

M. Collins. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 489–496, Philadephia, PA USA.

J.R. Curran and S. Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL)*, pages 164–167, Edmonton, Canada.

M. Dickinson and W.D. Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 107–114, Budapest, Hungary.

P. Ginsparg. 2001. Creating a global knowledge network. In *UNESCO Expert Conference on Electronic Publishing in Science*, Paris, France.

J. Grimm. 2003. Tralics, a LATEX to XML translator. *TUGboat*, 24(3):377 – 388.

B. Hachey, B. Alex, and M. Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Natural Language Learning (CoNLL)*, pages 144–151, Ann Arbor, MI USA.

R. J. Hanisch and P. J. Quinn. 2005. The IVOA. `http://www.ivoa.net/pub/info/`.

S. Harabagiu, D. Moldovan, M. Paşca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Gîrju, V. Rus, and P. Morărescu. 2000. Falcon: Boosting knowledge for answer engines. In *Proceedings of TREC-9*.

L. Hirschman and R. Gaizauskas. 2001. Natural language question answering: The view from here. *Journal of Natural Language Engineering*, 7(4):275–300.

L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu. 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(s1):i180–i182.

M.-C. Lortet, S. Borde, and F. Ochsenbein. 1994. Second Reference Dictionary of the Nomenclature of Celestial Objects. *A&AS*, 107:193–218, October.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

A. P. Martinez, S. Derriere, N. Gray, R. Mann, J. McDowell, T. McGlynn, Ochsenbein F., P. Osuna, G. Rixon, and R. Williams. 2005. The UCD1+ controlled vocabulary Version 1.02.

A. Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, Philadelphia, PA USA.

J.C. Reynar and A. Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 16–19, Washington DC, USA.

Daniel Trinkle. 2002. Detex. `http://www.cs.purdue.edu/homes/trinkle/detex/`.

Unicode Consortium. 2005. *The Unicode Standard*. Addison-Wesley, 4th edition.