From BioNER to AstroNER: Porting Named Entity Recognition to a New Domain

The SEER Project Team and Others

Edinburgh BioNER:	Bea Alex, Shipra Dingare, Claire Grover, Ben Hachey, Ewan Klein, Yuval Krymolowski, Malvina Nissim
Stanford BioNER:	Jenny Finkel, Chris Manning, Huy Nguyen
Edinburgh AstroNER:	Bea Alex, Markus Becker, Shipra Dingare, Rachel Dowsett, Claire Grover, Ben Hachey, Olivia Johnson, Ewan Klein, Yuval Krymolowski, Jochen Leidner, Bob Mann, Malvina Nissim, Bonnie Webber

Overview

- Named Entity Recognition
- The SEER project
- BioNER
- Porting to New Domains
- AstroNER

Named Entity Recognition

- As the first stage of Information Extraction, Named Entity Recognition (NER) identifies and labels strings in text as belonging to pre-defined classes of entities.
- (The second stage of Information Extraction (IE) identifies relations between entities.)
- NER or full IE can be useful technology for Text Mining.

Named Entity Recognition

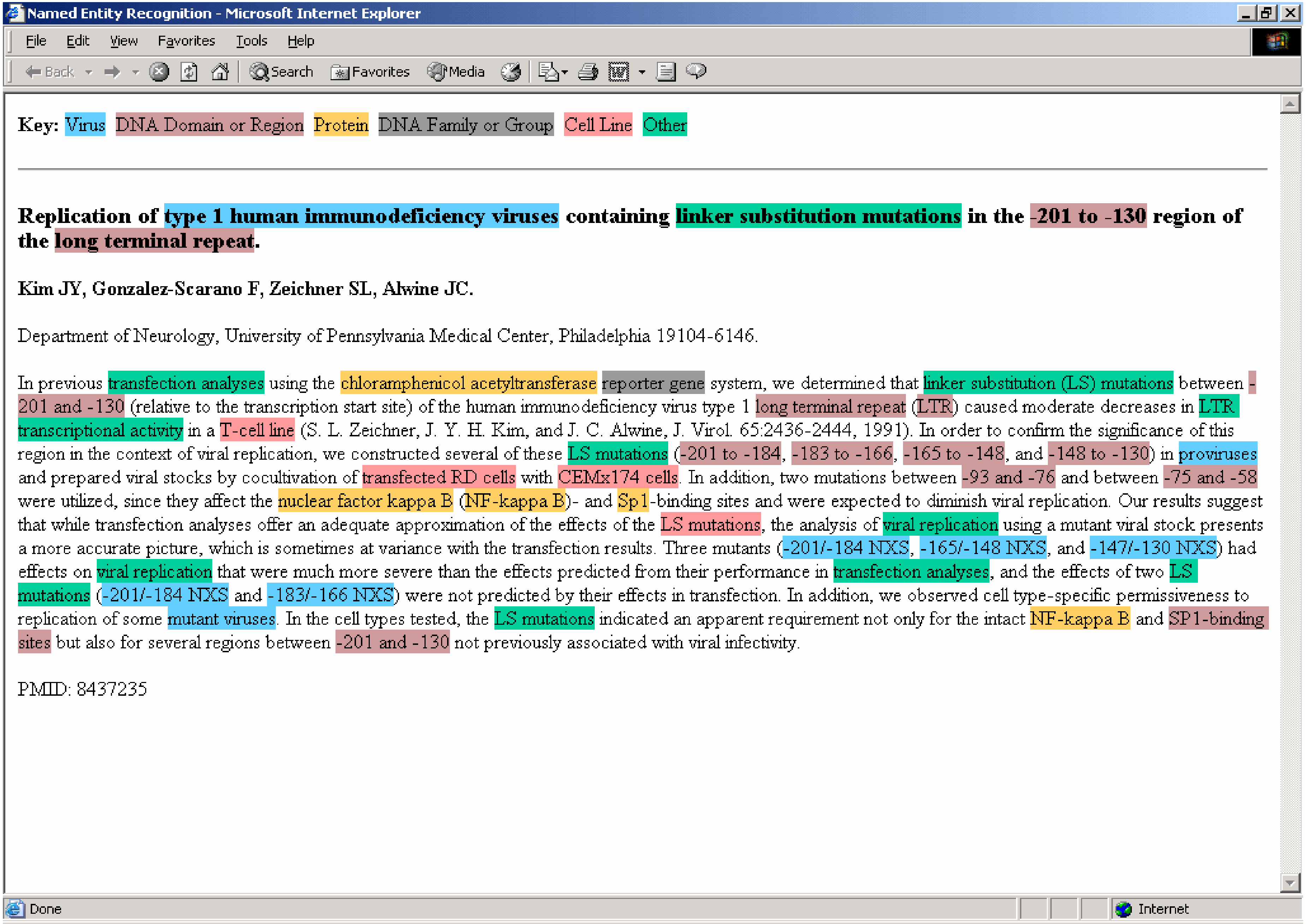
- Early work in NLP focused on general entities in newspaper texts e.g. person, organization, location, date, time, money, percentage

Newspaper Named Entities

Helen Weir, the finance director of Kingfisher, was handed a £334,607 allowance last year to cover the costs of a relocation that appears to have shortened her commute by around 15 miles. The payment to the 40-year-old amounts to roughly £23,000 a mile to allow her to move from Hampshire to Buckinghamshire after an internal promotion.

Named Entity Recognition

- For text mining from scientific texts, the entities are determined by the domain, e.g. for biomedical text, gene, virus, drug etc.


Named Entity Recognition - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Done
Internet

Key: Virus DNA Domain or Region Protein DNA Family or Group Cell Line Other

Replication of type 1 human immunodeficiency viruses containing linker substitution mutations in the -201 to -130 region of the long terminal repeat.

Kim JY, Gonzalez-Scarano F, Zeichner SL, Alwine JC.

Department of Neurology, University of Pennsylvania Medical Center, Philadelphia 19104-6146.

In previous transfection analyses using the chloramphenicol acetyltransferase reporter gene system, we determined that linker substitution (LS) mutations between -201 and -130 (relative to the transcription start site) of the human immunodeficiency virus type 1 long terminal repeat (LTR) caused moderate decreases in LTR transcriptional activity in a T-cell line (S. L. Zeichner, J. Y. H. Kim, and J. C. Alwine, J. Virol. 65:2436-2444, 1991). In order to confirm the significance of this region in the context of viral replication, we constructed several of these LS mutations (-201 to -184, -183 to -166, -165 to -148, and -148 to -130) in proviruses and prepared viral stocks by cocultivation of transfected RD cells with CEMx174 cells. In addition, two mutations between -93 and -76 and between -75 and -58 were utilized, since they affect the nuclear factor kappa B (NF-kappa B)- and Sp1-binding sites and were expected to diminish viral replication. Our results suggest that while transfection analyses offer an adequate approximation of the effects of the LS mutations, the analysis of viral replication using a mutant viral stock presents a more accurate picture, which is sometimes at variance with the transfection results. Three mutants (-201/-184 NXS, -165/-148 NXS, and -147/-130 NXS) had effects on viral replication that were much more severe than the effects predicted from their performance in transfection analyses, and the effects of two LS mutations (-201/-184 NXS and -183/-166 NXS) were not predicted by their effects in transfection. In addition, we observed cell type-specific permissiveness to replication of some mutant viruses. In the cell types tested, the LS mutations indicated an apparent requirement not only for the intact NF-kappa B and SP1-binding sites but also for several regions between -201 and -130 not previously associated with viral infectivity.

PMID: 8437235

The SEER Project

- Stanford-Edinburgh Entity Recognition
- Funded by the Edinburgh Stanford Link
  Jan 2002 — Dec 2004
- Focus:
  - NER technology applied in a range of new domains
  - generalise from named entities to include term entities
  - machine learning techniques in order to enable bootstrapping from small amounts of training data
- Domains: biomedicine, astronomy,
  archaeology

Biomedical NER Competitions

- BioCreative
  - Given a single sentence from a Medline abstract, identify all mentions of genes
  - “(or proteins where there is ambiguity)”
- BioNLP
  - Given full Medline abstracts, identify five types of entity
  - DNA, RNA, protein, cell line, cell type

The Biomedical NER Data

	Sentences	Words	NEs/Sent
	BioCreative		
Training	7,500	~200,000	~1.2
Development	2,500	~70,000	~1.2
Evaluation	5,000	~130,000	~1.2
	BioNLP		
Training	~19,000	~500,000	~2.75
Evaluation	~4,000	~100,000	~2.25

Evaluation Method

- Measure Precision, Recall and F-score.
- Both BioCreative and BioNLP used the exact-match scoring method
- Incorrect boundaries doubly penalized as false negatives and false positives.

chloramphenicol acetyl transferase reporter gene

chloramphenicol acetyl transferase reporter gene (FN)

transferase reporter gene (FP)

h

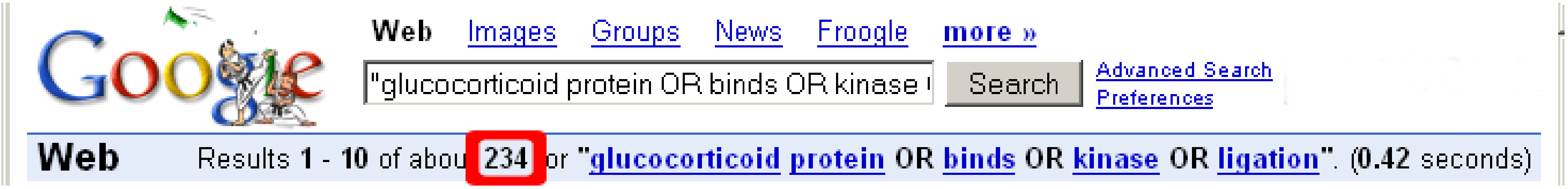
The SEER BioNER System

- Maximum Entropy Tagger in Java
  - Based on Klein et al (2003) CoNLL submission
  - Efforts mostly in finding new features
- Diverse Feature Set
  - Local Features
  - External Resources

External Resources

- Abbreviation
- TnT POS-tagger
- Frequency
- Gazetteers
- Web
- Syntax
- Abstract
- ABGENE/GENIA

Mining the Web



"glucocorticoid protein OR binds OR kinase

Web Results 1 - 10 of abou 234 or "glucocorticoid protein OR binds OR kinase OR ligation". (0.42 seconds)

Entity Type	Query	# hits
PROTEIN	"glucocorticoid protein OR binds OR kinase OR ligation"	*234*
DNA	"glucocorticoid dna OR sequence OR promoter OR site"	*101*
CELL_LINE	"glucocorticoid cells OR cell OR cell type OR line"	*1*
CELL_TYPE	"glucocorticoid proliferation OR clusters OR cultured OR cells"	*12*
RNA	"glucocorticoid mrna OR transcript OR	*35*

Feature Set

Word Features (All time s e.g. *Monday*, *April* are mapped to lower case)	w_i
	w_{i-1}
	w_{i+1}
	Last “real” word
	Next “real” word
	Any of the 4 previous words
	Any of the 4 next words
Bigrams	$w_i + w_{i-1}$
	$w_i + w_{i+1}$
TnT POS (trained on GENIA POS)	POS_i
	POS_{i-1}
	POS_{i+1}
Character Substrings	Up to a length of 6
Abbreviations	$abbr_i$
	$abbr_{i-1} + abbr_i$
	$abbr_i + abbr_{i+1}$
	$abbr_{i-1} + abbr_i + abbr_{i+1}$
Word + POS	$w_i + POS_i$
	$w_{i-1} + POS_i$
	$w_{i+1} + POS_i$

Word Shape	$shape_i$
	$shape_{i-1}$
	$shape_{i+1}$
	$shape_{i-1} + shape_i$
	$shape_i + shape_{i+1}$
	$shape_{i-1} + shape_i + shape_{i+1}$
Word Shape+Word	$w_{i-1} + shape_i$
	$w_{i+1} + shape_i$
Previous NE	NE_{i-1}
	$NE_{i-2} + NE_{i-1}$
	$NE_{i-1} + w_i$
Previous NE + POS	$NE_{i-1} + POS_{i-1} + POS_i$
	$NE_{i-2} + NE_{i-1} + POS_{i-2} + POS_{i-1} + POS_i$
Previous NE + Word Shape	$NE_{i-1} + shape_i$
	$NE_{i-1} + shape_{i+1}$
	$NE_{i-1} + shape_{i-1} + shape_i$
	$NE_{i-2} + NE_{i-1} + shape_{i-2} + shape_{i-1} + shape_i$
Parentheses	Paren-Matching – a feature that signals when one parentheses in a pair has been assigned a different tag than the other in a window of 4 s

Postprocessing – BioCreative

- Discarded results with mismatched parentheses
- Different boundaries were detected when searching the sentence forwards versus backwards
- Unioned the results of both; in cases where boundary disagreements meant that one detected gene was contained in the other, we kept the shorter gene

Results

BioCreative	Precision	Recall	F-Score
Gene/Protein	0.828	0.836	0.832

BioNLP	Precision	Recall	F-Score
Protein	0.774	0.685	0.727
DNA	0.662	0.696	0.679
RNA	0.720	0.659	0.688
Cell Line	0.590	0.471	0.524
Cell Type	0.626	0.770	0.691
Overall	0.716	0.686	0.701

What If You Lack Training Data?

- When porting to a new domain it is likely that there will be little or no annotated data available.
- Do you pay annotators to create it?
- Are there methods that will allow you to get by with just a small amount of data?
- Bootstrapping Techniques

AstroNER: The ‘Surprise’ Task

- Aims
  - simulate a practical situation
  - experiment with bootstrapping methods
  - gain practical experience in porting our technology to a new domain using limited resources
  - monitor resource expenditure to compare the practical utility of various methods
- Collaborators: Bonnie Webber,
  Bob Mann

Method

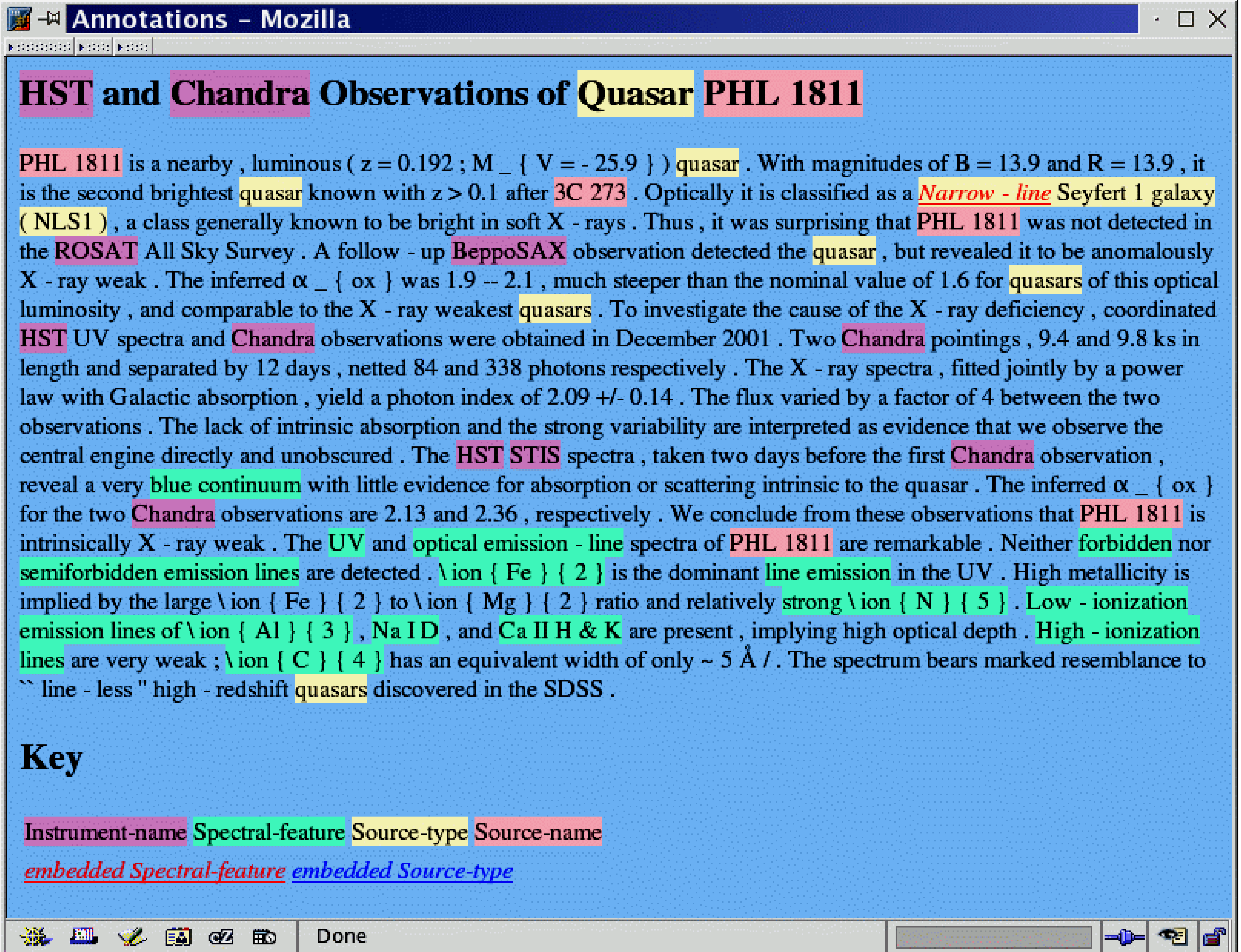
- The data was chosen and prepared in secret to ensure fair comparison.
- The training set was kept very small but large amounts of tokenised unlabelled data were made available.
- Three teams, each given the same period of time to perform task
- Approaches:
  - co-training, weakly supervised learning, active learning

Data and Annotation

- Astronomy abstracts from the NASA Astrophysics Data Service (http://adsabs.harvard.edu/) 1997-2003.
- Sub-domain: spectroscopy/spectral lines
- 4 entity types: instrument-name, spectral-feature, source-type, source-name
- Data:

	abstracts	sentences	entities
training	50	502	874
testing	159	1,451	2,568
unlabeled	778	7,979	

- Annotation tool based on the NXT toolkit for expert annotation of training & testing sets as well as active learning annotation.



HST and Chandra Observations of Quasar PHL 1811

PHL 1811 is a nearby , luminous (z = 0.192 ; M _ { V = - 25.9 }) quasar . With magnitudes of B = 13.9 and R = 13.9 , it is the second brightest quasar known with z > 0.1 after 3C 273 . Optically it is classified as a *Narrow - line* Seyfert 1 galaxy (NLS1) , a class generally known to be bright in soft X - rays . Thus , it was surprising that PHL 1811 was not detected in the ROSAT All Sky Survey . A follow - up BeppoSAX observation detected the quasar , but revealed it to be anomalously X - ray weak . The inferred α _ { ox } was 1.9 -- 2.1 , much steeper than the nominal value of 1.6 for quasars of this optical luminosity , and comparable to the X - ray weakest quasars . To investigate the cause of the X - ray deficiency , coordinated HST UV spectra and Chandra observations were obtained in December 2001 . Two Chandra pointings , 9.4 and 9.8 ks in length and separated by 12 days , netted 84 and 338 photons respectively . The X - ray spectra , fitted jointly by a power law with Galactic absorption , yield a photon index of 2.09 +/- 0.14 . The flux varied by a factor of 4 between the two observations . The lack of intrinsic absorption and the strong variability are interpreted as evidence that we observe the central engine directly and unobscured . The HST STIS spectra , taken two days before the first Chandra observation , reveal a very blue continuum with little evidence for absorption or scattering intrinsic to the quasar . The inferred α _ { ox } for the two Chandra observations are 2.13 and 2.36 , respectively . We conclude from these observations that PHL 1811 is intrinsically X - ray weak . The UV and optical emission - line spectra of PHL 1811 are remarkable . Neither forbidden nor semiforbidden emission lines are detected . \ ion { Fe } { 2 } is the dominant line emission in the UV . High metallicity is implied by the large \ ion { Fe } { 2 } to \ ion { Mg } { 2 } ratio and relatively strong \ ion { N } { 5 } . Low - ionization emission lines of \ ion { Al } { 3 } , Na I D , and Ca II H & K are present , implying high optical depth . High - ionization lines are very weak ; \ ion { C } { 4 } has an equivalent width of only ~ 5 Å / . The spectrum bears marked resemblance to `` line - less '' high - redshift quasars discovered in the SDSS .

Key

Instrument-name Spectral-feature Source-type Source-name

embedded Spectral-feature *embedded Source-type*

Co-training

- Basic idea: use the strengths of one classifier to rectify the weaknesses of another.
- Two different methods classify a set of seed data; select results of one iteration, and add them to the training data for the next iteration.
- Various choices:
  - same classifier with different feature splits, or two different classifiers
  - cache size (# examples to tag on each iteration)
  - add labeled data to new training set if both agree, or add labeled data from one to training set of the other
  - retrain some or all classifiers at each iteration

		#sentences	#words	#entities	#classes
SEED	bio-data	500	12,900	1,545	5+1
	astro-data	502	15,429	874	4+1
TEST	bio-data	3,856	101039	8,662	5+1
	astro-data	1,451	238,655	2,568	4+1

UNLABELLED DATA: ca. 8,000 sentences for both sets

START PERFORMANCE (F)	Stanford	C&C	TnT	YAMCHA
bio-data	56.87	48.42	41.62	50.64
astro-data	69.06	64.47	61.45	61.98

- best settings on biomedical data:
    - Stanford, C&C, and TnT; cache=200; agreement; retrain Stanford only
    - Stanford and YAMCHA; cache=500; agreement
    - NOTE: in both cases limited improvement (max 2 percentage points)
- on astronomical data: no real positive results so far

TAKE HOME MESSAGE: COTRAINING QUITE UNSUCCESFUL FOR THIS TASK!

REASONS: Classifiers not different enough? Classifiers not good enough to start with?

Weakly supervised

- Many multi-token entities, typically a head word preceded by modifiers:
  - instrument-name: `Very Large` **`Telescope`**
  - source-type: `radio–quiet` **`QSOs`**
  - spectral-feature: `[O II]` **`emission`**
- Find most likely modifier sequences for a given initial set of concepts
- Build a gazetteer for each entity subtype and use it for markup.
- Results: F-score = 49%.

Active Learning

- Supervised Learning
  - Select random examples for labeling
  - Requires large amount of (relatively expensive) annotated data
- Active Learning
  - Select most ‘informative’ examples for labelling
  - Maximal reduction of error rate with minimal amount of labelling
  - Faster converging learning curves
    - Higher accuracy for same amount of labelled data
    - Less labelled data for same levels of accuracy

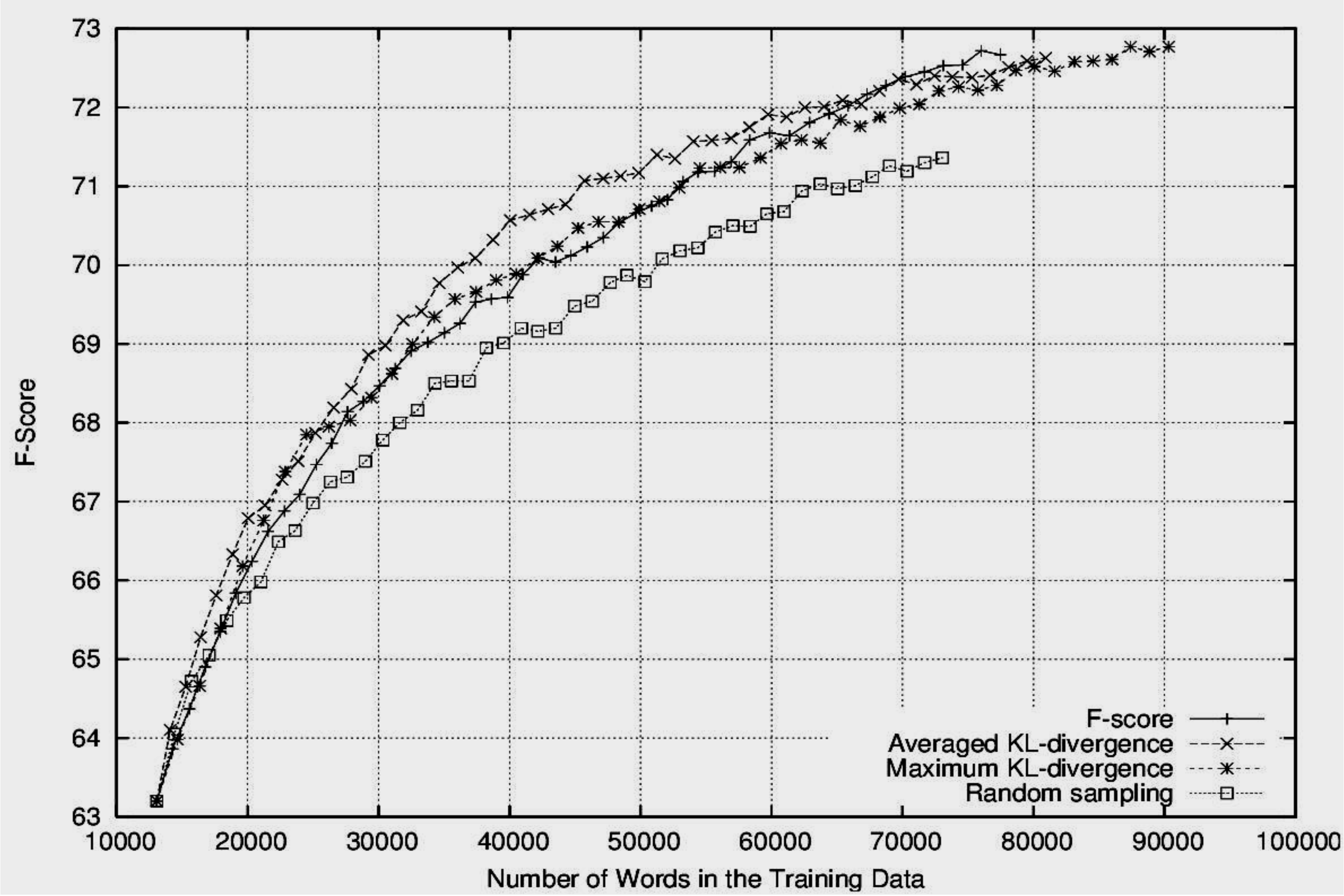
Parameters

- Annotation level: Document? Sentence? Word?
- Selection method:
  - Query-by-committee with several sample selection metrics
    - Average KL-divergence
    - Maximum KL-divergence
    - F-score
- Batch size: 1 ideal but impractical. 10? 50? 100?

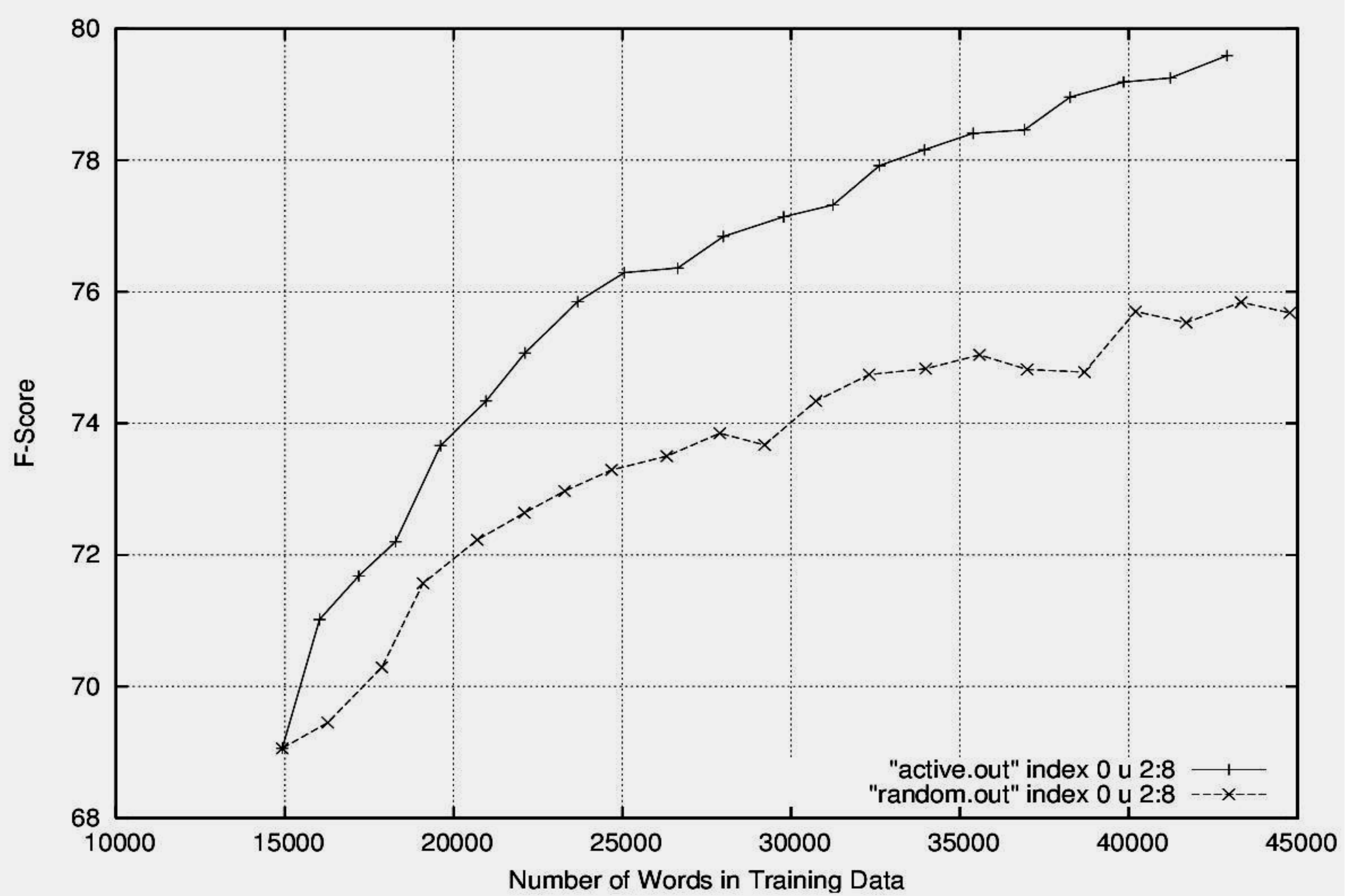
Experiments

- BioNLP
  - Corpus: developed for BioNLP 2004 shared task, based on GENIA corpus
  - Entities: DNA, RNA, cell-line, cell-type, protein
  - Experiments: 10 fold cross validation used to tune AL parameters for real experiments
- AstroNER
  - Experiments: 20 rounds of annotation with active sample selection

BioNLP: Words vs. F-score



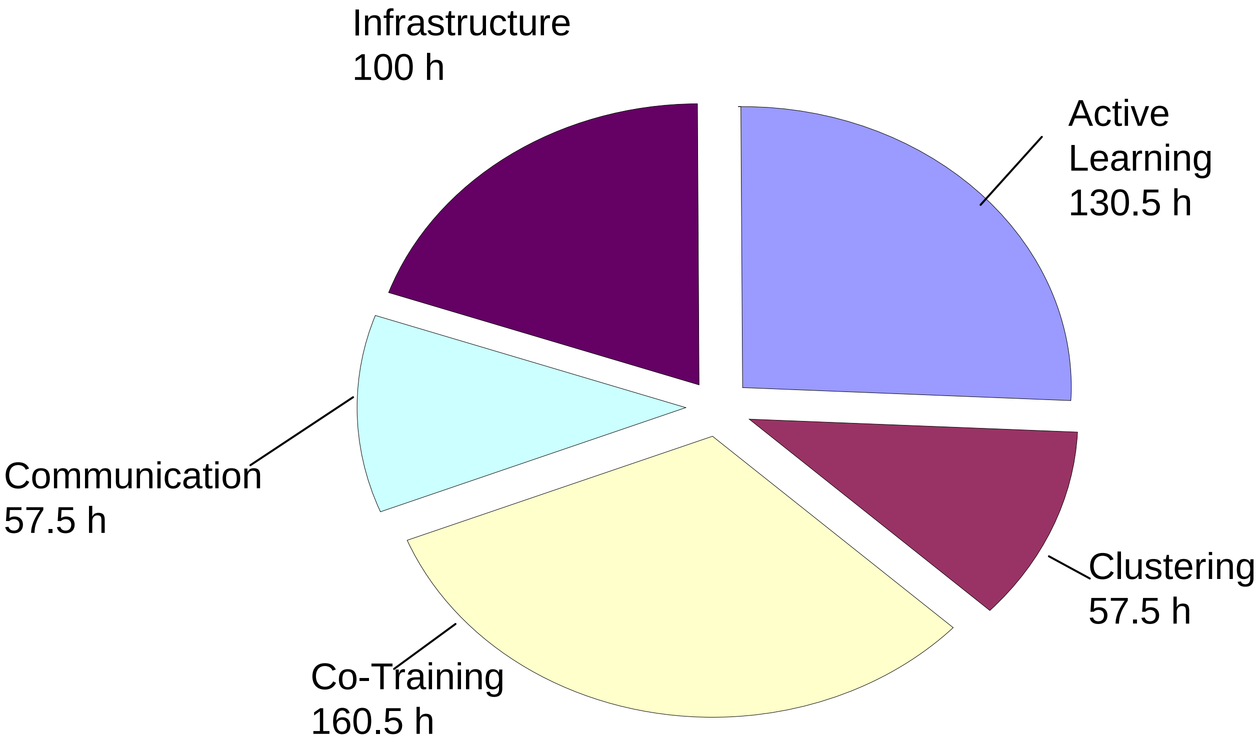
AstroNER: Words vs. F-score



Time Monitoring

- Objective:
  - Progress towards NL engineering (cost/time-aware)
- Method:
  - Web-based time tracking tool used to record how time was spent
  - Separation between shared (communication, infastructure) and method-specific time use
- Result:
  - No dramatic cost differences between 3 methods
  - Roughly 64 person days total cost (all methods)

Time Monitoring


Infrastructure
100 h
Active
Learning
130.5 h
Communication
57.5 h
Clustering
57.5 h
Co-Training
160.5 h