Parsing the Language of Web 2.0

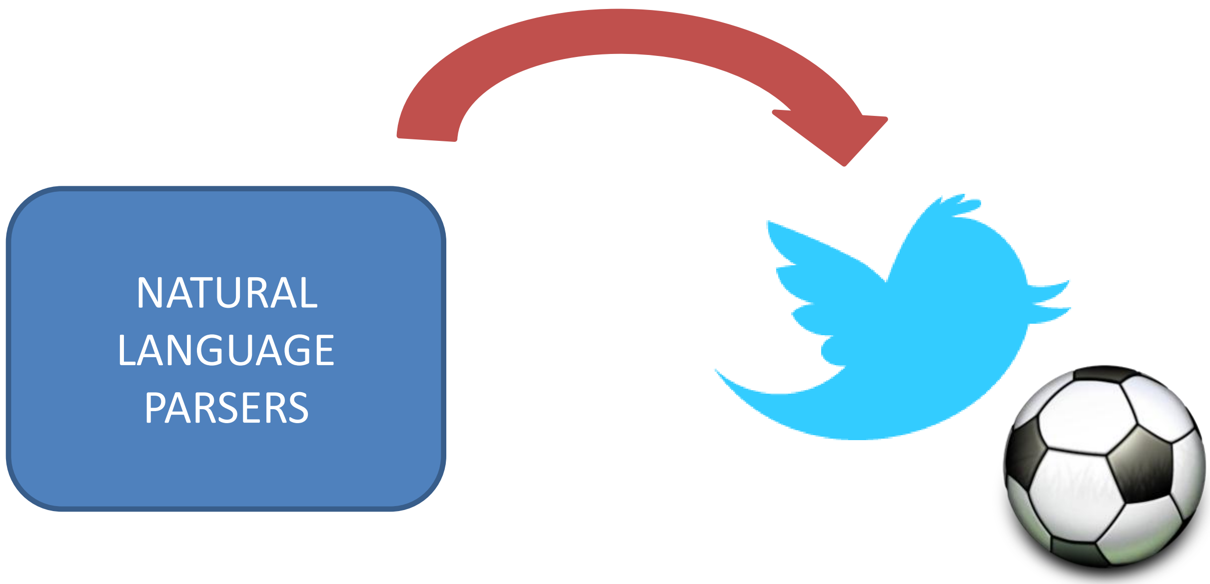
Jennifer Foster

Joint work with Joachim Wagner, Özlem Çetinoğlu, Joseph Le Roux, Joakim Nivre, Anton Bryl, Rasul Kaljahi, Johann Roturier, Deirdre Hogan, Raphael Rubino, Fred Hollowood and Josef van Genabith

Oslo, May 9th 2012

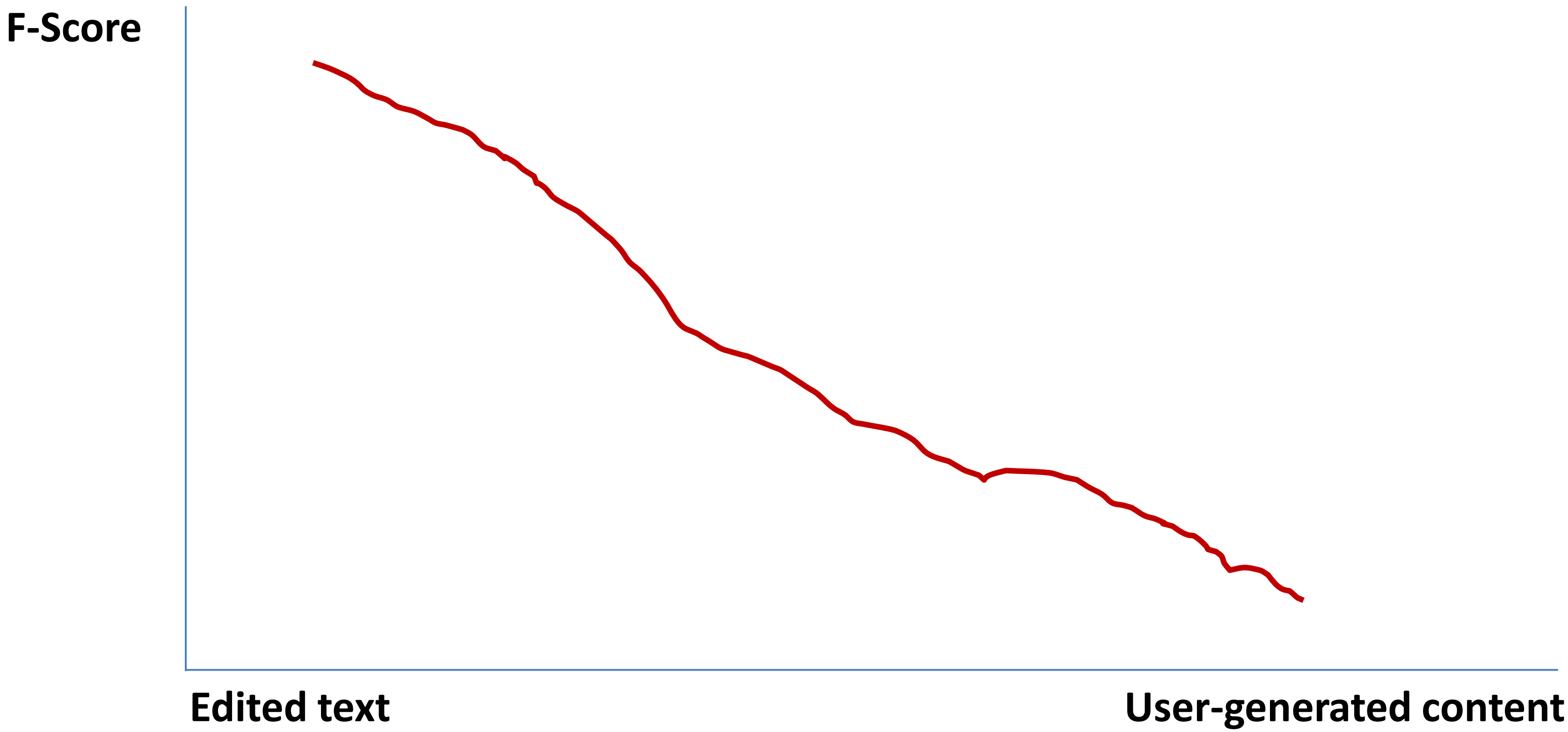
What are we doing?

1. Apply off-the-shelf part-of-speech taggers and syntactic parsers to the language of social media



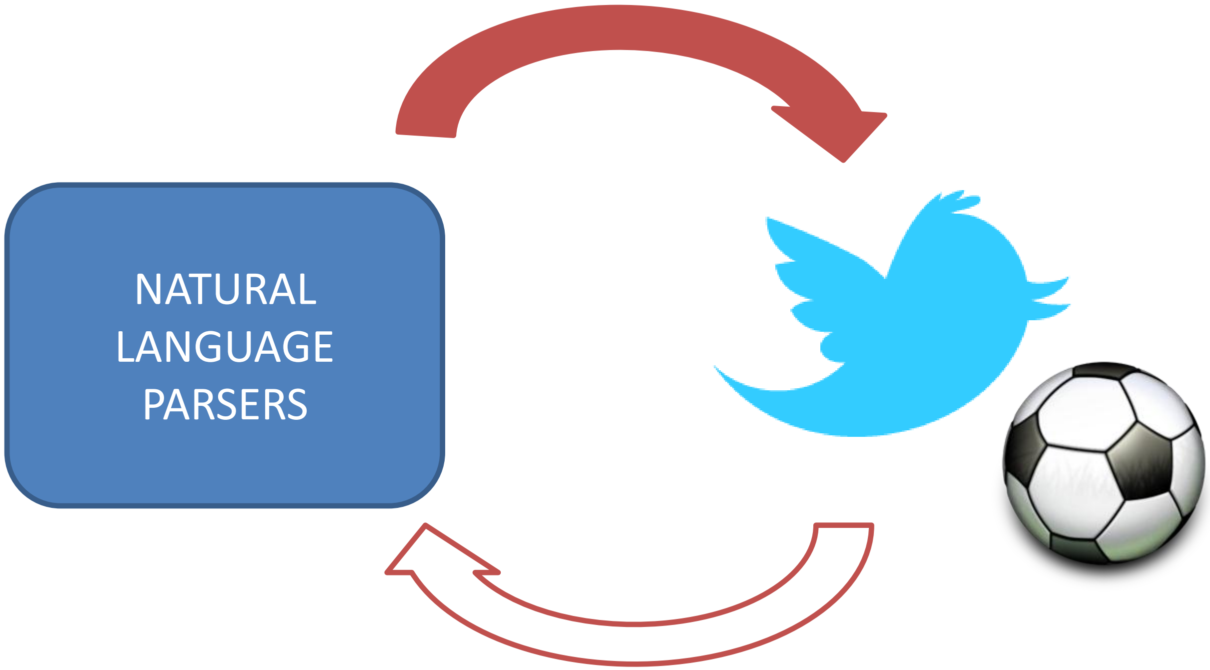
What are we doing?

2. Investigate the drop in performance



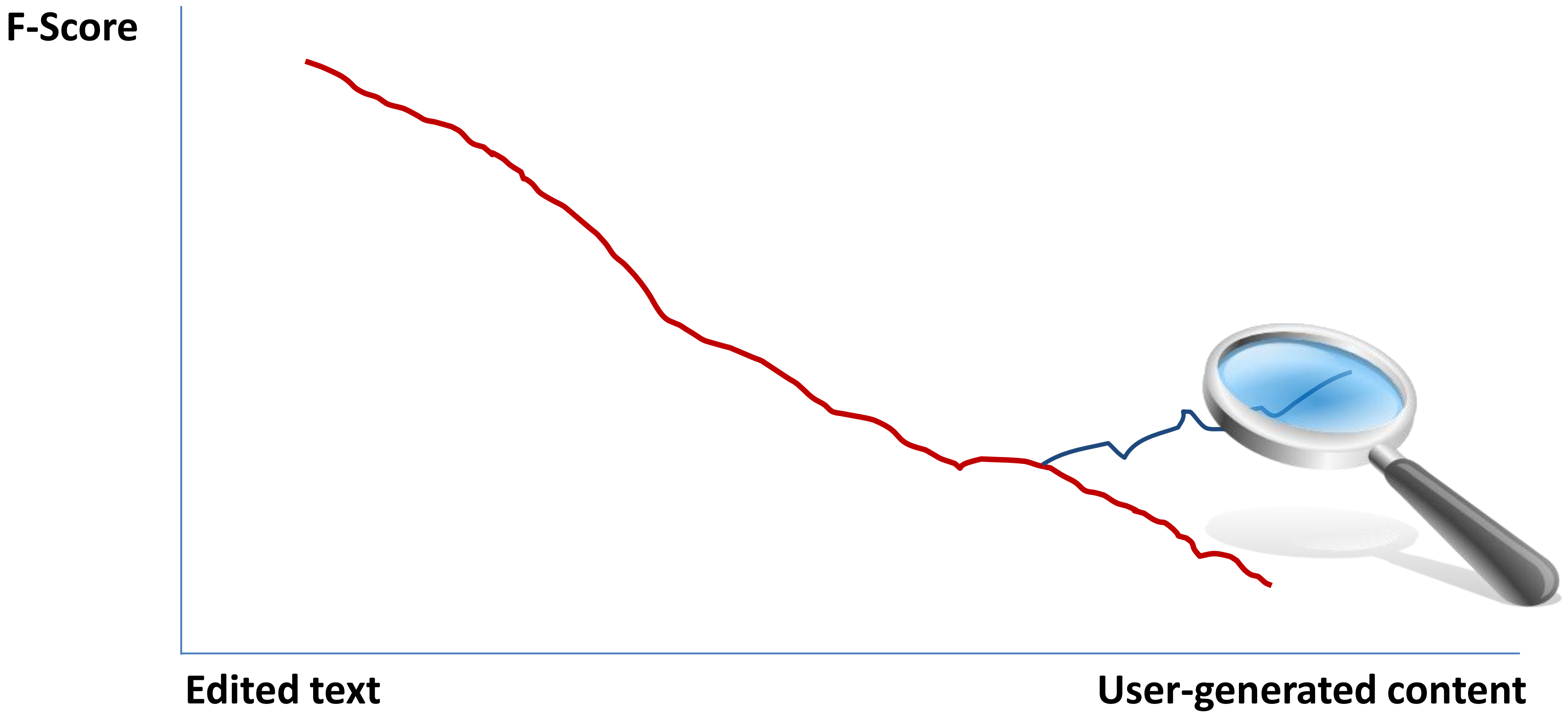
What are we doing?

3. Retrain tools on automatically analysed Web2.0 data

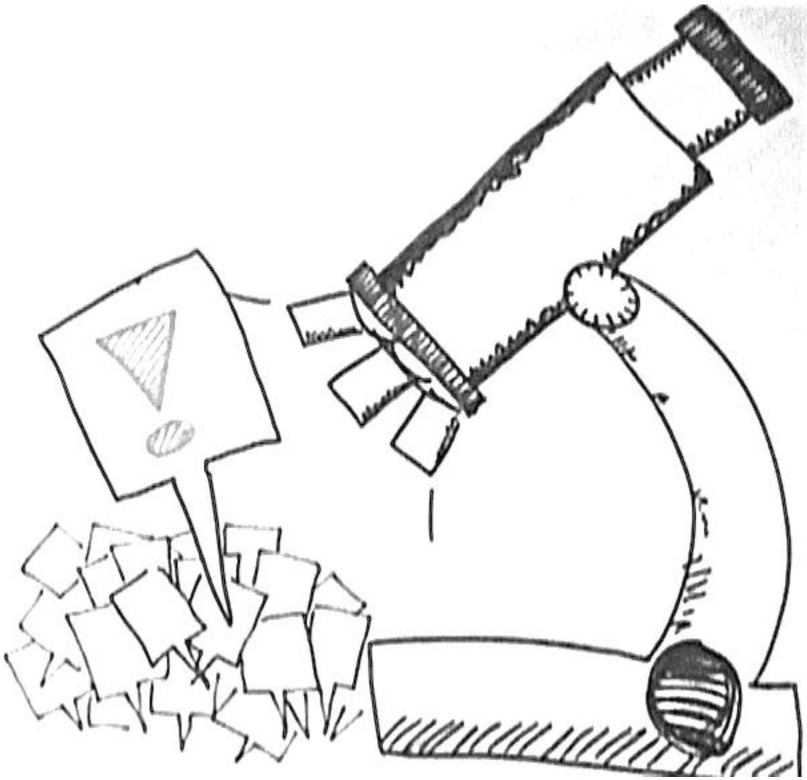


What are we doing?

4. Investigate the changes



Why parsing?



- Assign structure to text.
- Who did what to whom?
- Useful for various «sense-making» applications
- MT, QA, Sentiment Analysis

Why the language of Web 2.0?

- Explosive growth in social media
- Cultural and commercial interest

Why is this a challenge?

- WSJ-trained statistical parsers perform very well on edited text
  - Not designed to work on noisy, unedited language

Why is this a challenge?

- WSJ-trained statistical parsers perform very well on edited text
  - Not designed to work on noisy, unedited language
- Can standard domain adaptation techniques be applied?

Why is this a challenge?

- WSJ-trained statistical parsers perform very well on edited text
  - Not designed to work on noisy, unedited language
- Can standard domain adaptation techniques be applied?
- Potential obstacles:
  - Not enough labelled data
  - Web2.0 is not really a domain

Talk Structure

1. Pilot Study (Foster 2010)

Talk Structure

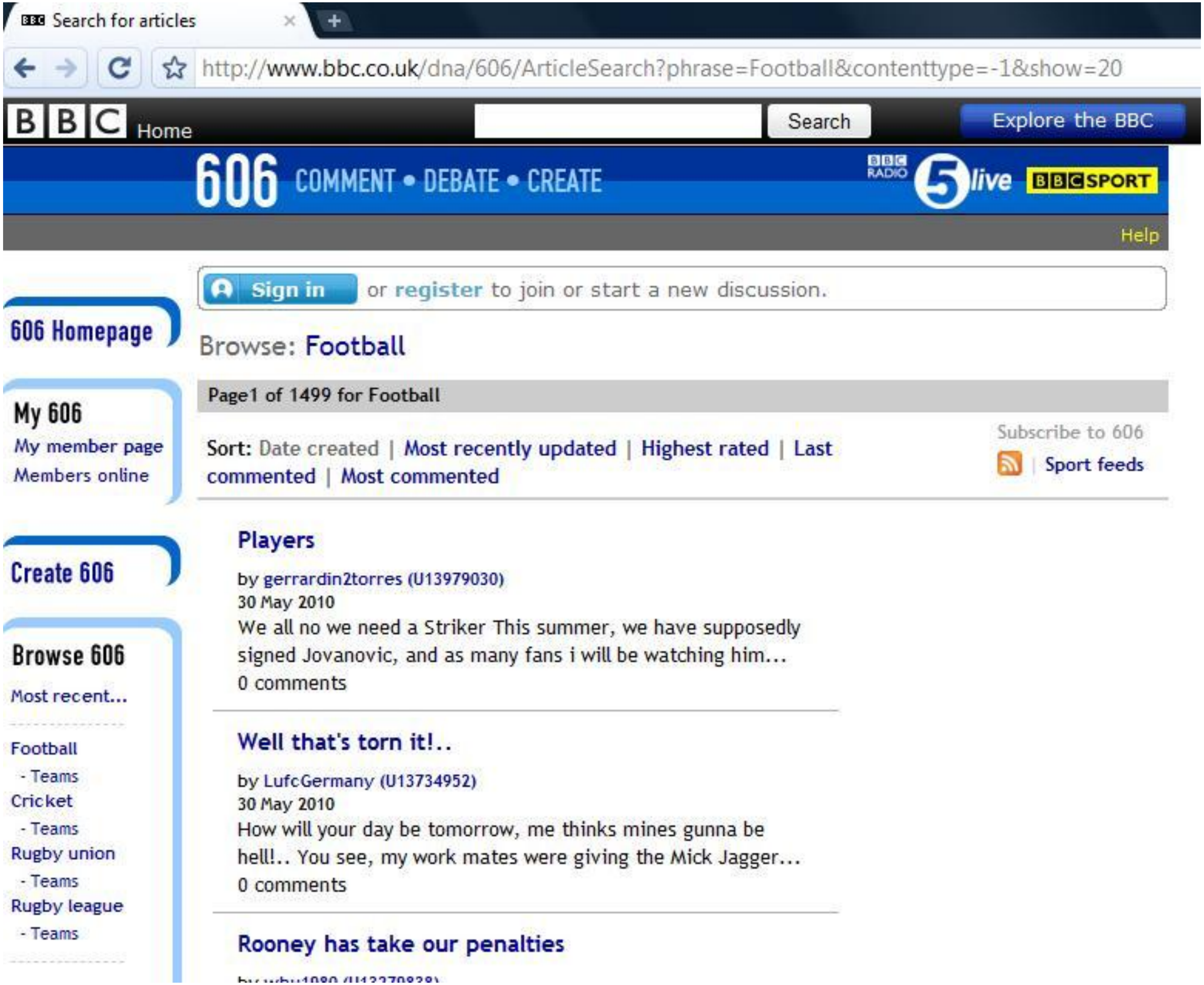
1. Pilot Study (Foster 2010)
2. More data, more parsers, more experiments (Foster et al. 2011)

Talk Structure

1. Pilot Study (Foster 2010)
2. More data, more parsers, more experiments (Foster et al. 2011)
3. Current Work:
   – SANCL Shared Task on Parsing Web Data
   – Confident MT Project

Part One

Pilot Study


Search for articles

http://www.bbc.co.uk/dna/606/ArticleSearch?phrase=Football&contenttype=-1&show=20

BBC Home

Search

Explore the BBC

606 COMMENT • DEBATE • CREATE

BBC RADIO 5 live BBC SPORT

Help

606 Homepage

My 606

My member page

Members online

Create 606

Browse 606

Most recent...

Football

- Teams

Cricket

- Teams

Rugby union

- Teams

Rugby league

- Teams

Sign in or register to join or start a new discussion.

Browse: Football

Page1 of 1499 for Football

Sort: Date created | Most recently updated | Highest rated | Last commented | Most commented

Subscribe to 606

Sport feeds

Players

by gerrardin2torres (U13979030)
30 May 2010
We all no we need a Striker This summer, we have supposedly signed Jovanovic, and as many fans i will be watching him...
0 comments

Well that's torn it!..

by LufcGermany (U13734952)
30 May 2010
How will your day be tomorrow, me thinks mines gunna be hell!.. You see, my work mates were giving the Mick Jagger...
0 comments

Rooney has take our penalties

Forum Data Examples

If anything is going to happen to change how the game is controlled on the pitch, Sir Alex and other persistent whingers like Steve Bruce and Arsene Wenger need to crititque the refereeing from a whole game perpsective, not just the incidents they see through their red tinted spectacles. How refreshing that would be.

Forum Data Examples

If anything is going to happen to change how the game is controlled on the pitch, Sir Alex and other persistent whingers like Steve Bruce and Arsene Wenger need to crititque the refereeing from a whole game perpsective, not just the incidents they see through their red tinted spectacles. How refreshing that would be.

Forum Data Examples

havent man c got a good team now if thay ceep geting grate players all of there normal players will lose out for instans thay got given so joe hart hat to go on lone to bermingham !!!!! and thats just one player how was left out

Forum Data Examples

- *He overpowered the guy*
- *He didn't.*
- *Where was drogba yesterday?*

Forum Data Examples

- *Try again fella (going to school that is)*

Forum Data Examples

- *Try again fella (going to school that is)*
- *Why are most the posts on here like essays?*

Forum Data Examples

- *Try again fella (going to school that is)*
- *Why are most the posts on here like essays?*
- *your lose to Wigan and Bolton would be more scrutunized (cba to check spelling) than it has been this year.*

Dataset

Development set

- 42 posts
- 185 sentences
- On average, 18 words per sentence

Dataset

Development set

- 42 posts
- 185 sentences
- On average, 18 words per sentence

Test Set

- 40 posts
- 170 sentences
- On average, 15 words per sentence

Annotation Process

- Manual tokenisation and spell correction
- Parse trees produced by the Bikel parser corrected by hand
- Penn Treebank bracketing guidelines
- Function tags and traces not annotated
- Difficult decisions were documented
- Two passes through the data

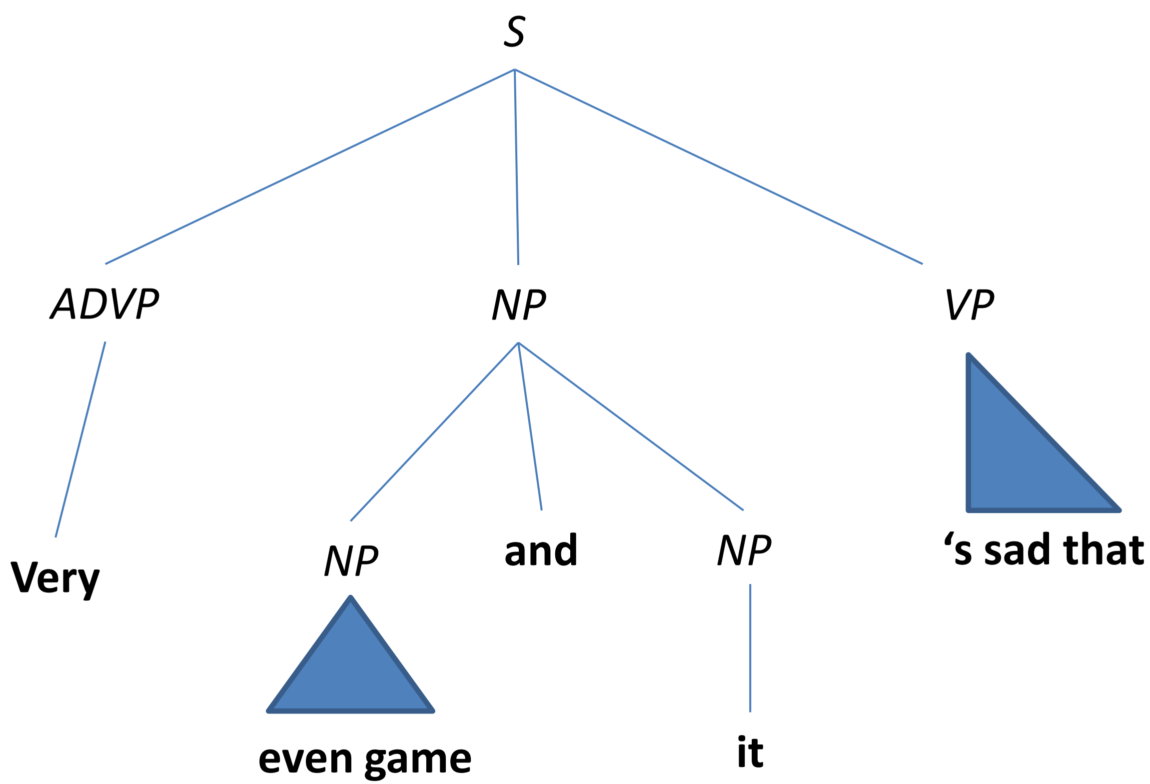
Parser Evaluation

Performance of Berkeley parser (Petrov et al. 2006)

Test Set	Recall	Precision	F-Score
WSJ23	88.88	89.46	89.17
Football Gold Tokens+Spell	78.15	76.97	77.56

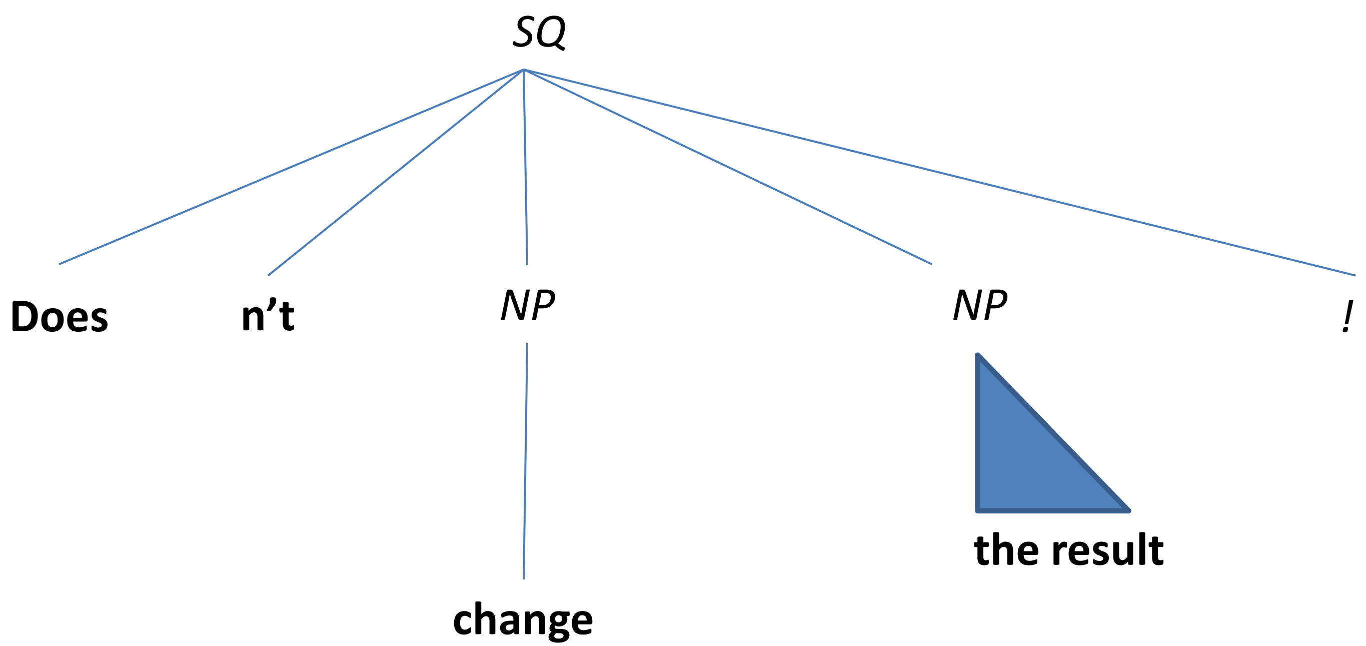
Unlike Constituent Coordination

Very even game and it's sad that….



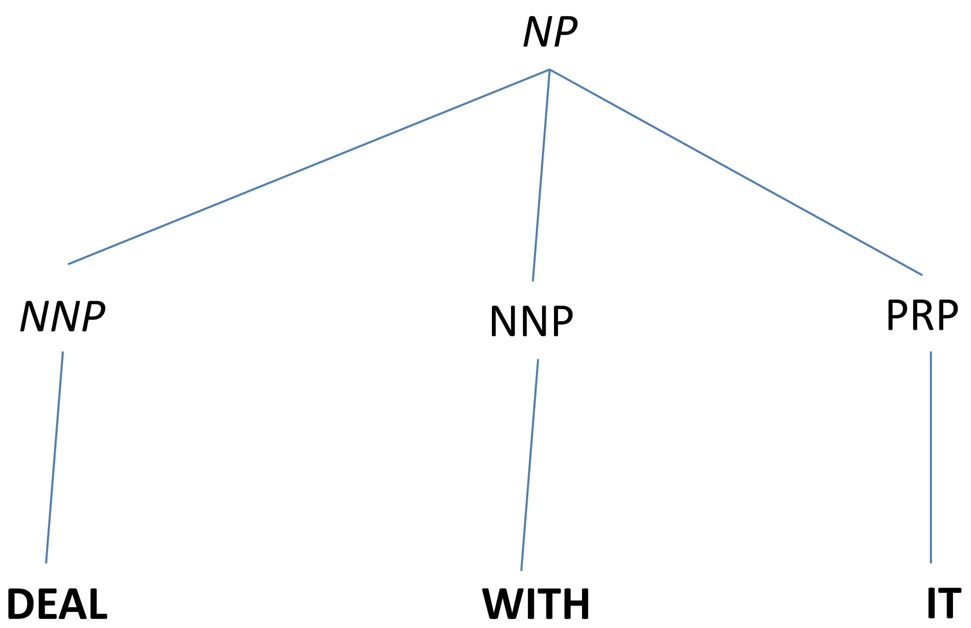
Subject Ellipsis

Does n't change the result !



Non-standard capitalisation

DEAL WITH IT



Qualitative Evaluation

- Unlike constituent coordination
- Subject ellipsis
- Stream-of-consciousness sentence coordination
- Abbreviations and acronyms
- Domain-specific idioms
- Non-standard capitalisation
- Lack of apostrophes
- Function word misspelling

Part Two

More data, more parsers, more experiments

i heart beltran

On Fox: RNC chair sends letter to GOP calling Obama "ARROGANT" " #tcot #sgp #hhrs

Twas okay.

FF > S4

Very even game and it's sad that....

Doesn't change the result though.

I just think he looks like a big baby , and ppl USED to call him that

LOL!

or it was cos you lost

i heart beltran

On Fox: RNC chair sends letter to GOP calling Obama “ARROGANT” ” #tcot #sgp #hhrs

Twas okay.

FF > S4

Very even game and it’s sad that….

Doesn’t change the result though.

I just think he looks like a big baby , and ppl USED to call him that

LOL!

or it was cos you lost

i heart beltran

On Fox: RNC chair sends letter to GOP calling Obama "ARROGANT" " #tcot #sgp #hhrs

Twas okay.

FF > S4

Very even game and it's sad that….

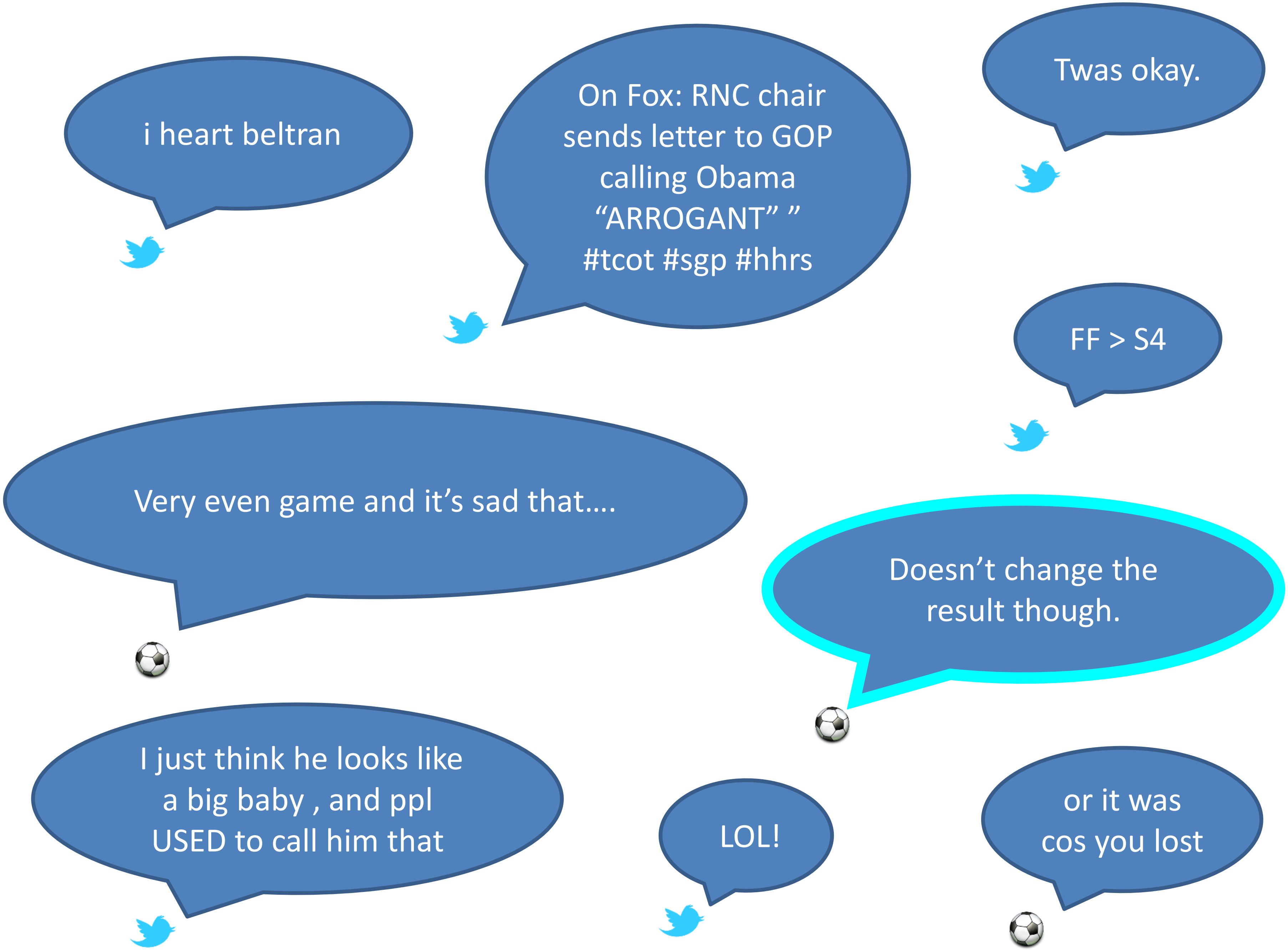
Doesn't change the result though.

I just think he looks like a big baby , and ppl USED to call him that

LOL!

or it was cos you lost


i heart beltran
On Fox: RNC chair sends letter to GOP calling Obama "ARROGANT" " #tcot #sgp #hhrs
Twas okay.
FF > S4
Very even game and it's sad that….
Doesn't change the result though.
I just think he looks like a big baby , and ppl USED to call him that
LOL!
or it was cos you lost


i heart beltran
On Fox: RNC chair sends letter to GOP calling Obama "ARROGANT" " #tcot #sgp #hhrs
Twas okay.
FF > S4
Very even game and it's sad that....
Doesn't change the result though.
I just think he looks like a big baby , and ppl USED to call him that
LOL!
or it was cos you lost

i heart beltran

On Fox: RNC chair sends letter to GOP calling Obama “ARROGANT” ” #tcot #sgp #hhrs

Twas okay.

FF > S4

Very even game and it’s sad that….

Doesn’t change the result though.

I just think he looks like a big baby , and ppl USED to call him that

LOL!

or it was cos you lost


i heart beltran
On Fox: RNC chair sends letter to GOP calling Obama “ARROGANT” ” #tcot #sgp #hhrs
Twas okay.
FF > S4
Very even game and it’s sad that….
Doesn’t change the result though.
I just think he looks like a big baby , and ppl USED to call him that
LOL!
or it was cos you lost

Datasets

Corpus Name	#Sentences	Average Sent. Length	Median Sent. Length	Std. Deviation
TwitterDev	269	11.1	10	6.4
TwitterTest	250	11.3	10	6.8
TwitterTrain	1.4 million	8.6	7	6.1
FootballDev	258	17.7	14	13.9
FootballTest	223	16.1	14	9.7
FootballTrain	1 million	15.4	12	13.3

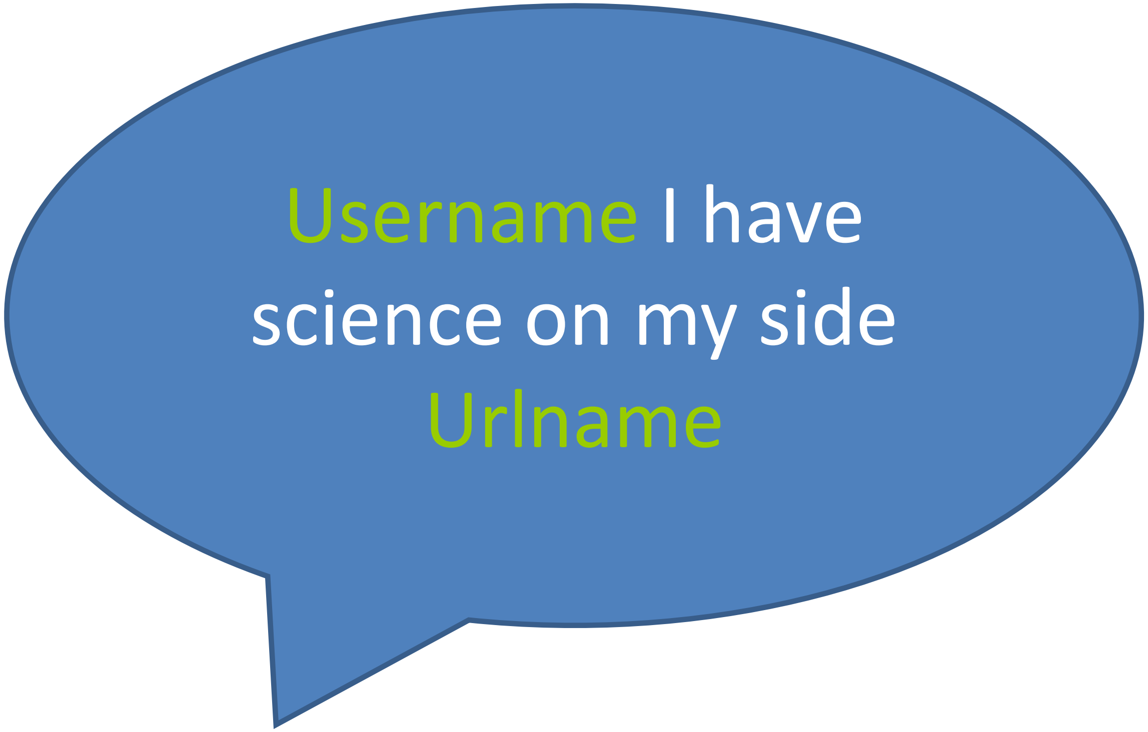
Pre-processing

@joebloggs I have science on my side http://bit.ly/gV4iUH

Pre-processing

Username I have
science on my side
Urlname

Pre-processing



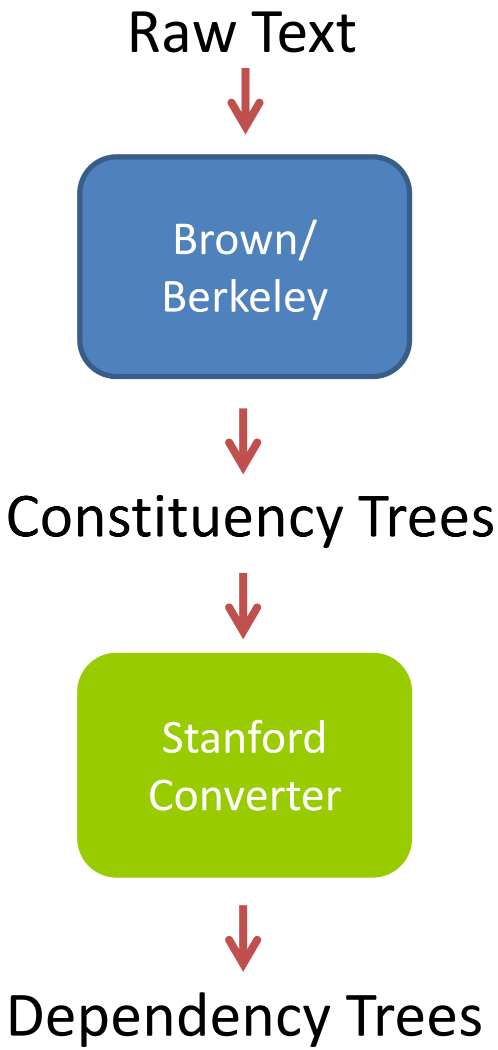
Transformations applied to both training *and* test/dev data.

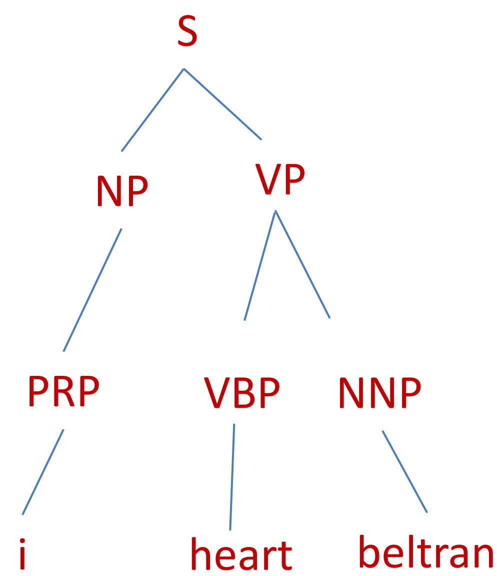
Pre-processing

Difference between training and test/dev data:

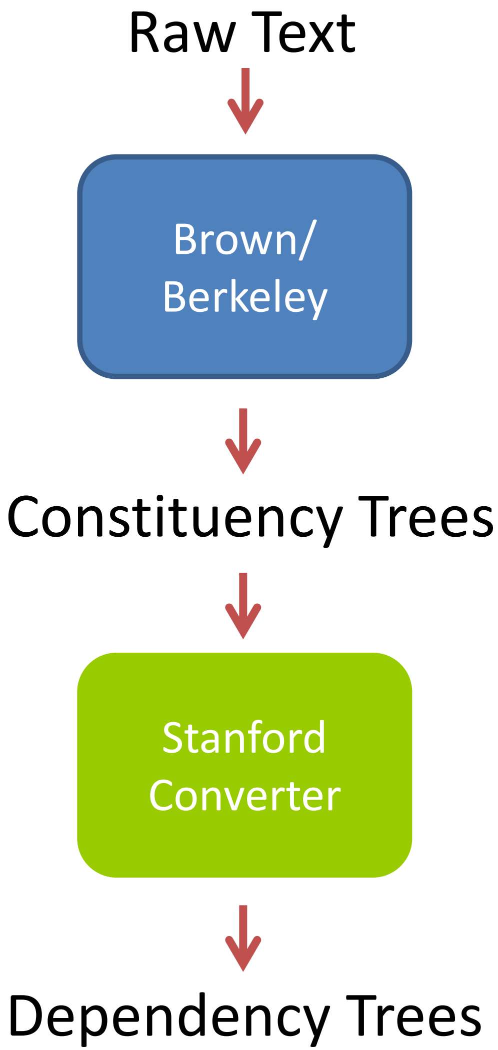
- Training data is split into sentences and tokenised *automatically.*
- Test/dev data is split into sentences and tokenised *manually* before syntactic annotation.

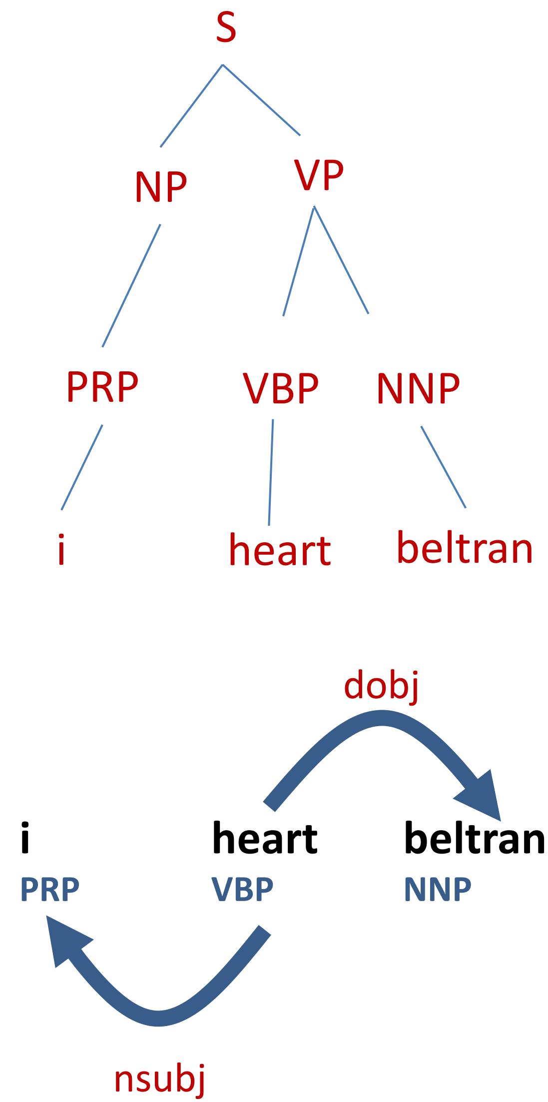
Baseline Models - Constituency



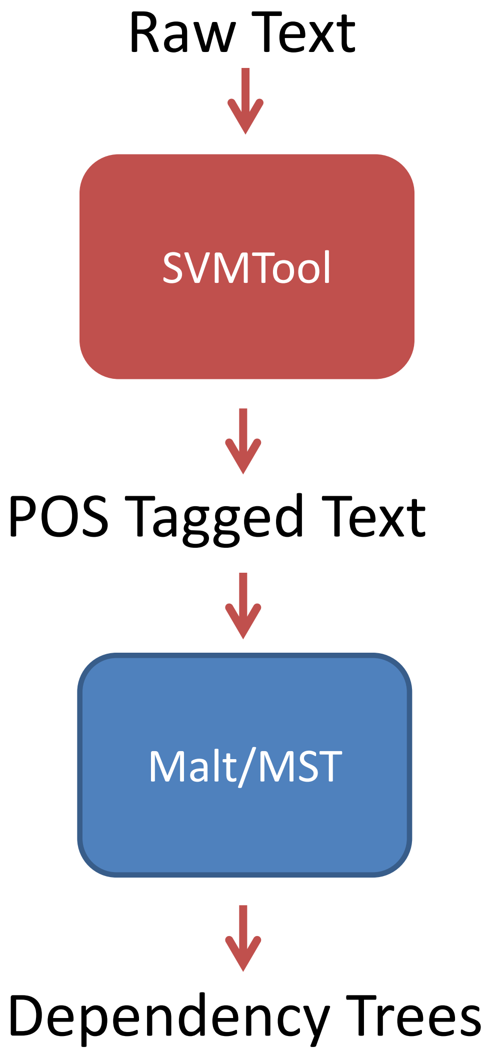


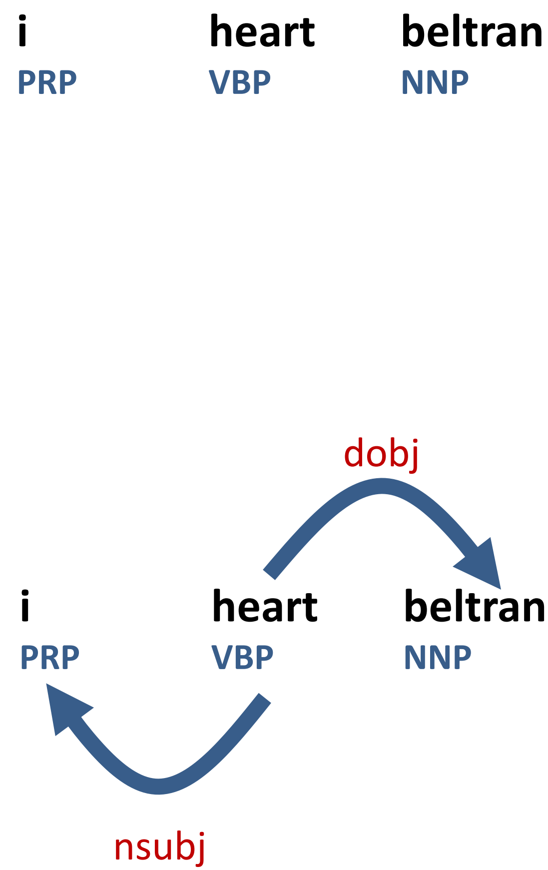
Baseline Models - Constituency





Baseline Models - Dependency





Baseline Results – Constituency

- F-scores:

WSJ22	*FootballDev*	*TwitterDev*
89 - 91.9	78.8 - 79.7	70.1 - 73.8

Baseline Results – Constituency

- F-scores:

WSJ22	*FootballDev*	*TwitterDev*
89 - 91.9	78.8 - 79.7	70.1 - 73.8

- Brown > Berkeley own POS > Berkeley predicted POS

Baseline Results – Constituency

- F-scores:

WSJ22	*FootballDev*	*TwitterDev*
89 - 91.9	78.8 - 79.7	70.1 - 73.8

- Brown > Berkeley own POS > Berkeley predicted POS
- Twitter data is harder to parse than the discussion forum data

Baseline Results - Dependency

- LAS:

WSJ22	*FootballDev*	*TwitterDev*
88 - 91.5	76.4 - 82	67.3 - 71.4

Baseline Results - Dependency

- LAS:

WSJ22	*FootballDev*	*TwitterDev*
88 - 91.5	76.4 - 82	67.3 - 71.4

- Brown > Berkeley own/predicted POS > MST > Malt

Baseline Results – POS Tagging

- POS Tagging Accuracy

WSJ22	*FootballDev*	*TwitterDev*
96.3 - 96.6	92.2 - 93.5	84.1- 85.5

Baseline Results – POS Tagging

- POS Tagging Accuracy

WSJ22	*FootballDev*	*TwitterDev*
96.3 - 96.6	92.2 - 93.5	84.1- 85.5

- Unknown Word Rate

WSJ22	*FootballDev*	*TwitterDev*
2.8%	6.8%	16.6%

POS Tagging and Parsing

- Effect of Gold POS Tagging on LAS

WSJ22	*FootballDev*	*TwitterDev*
+ 1.1 - 2.0	+ 3.0 - 4.4	+ 7.9 - 11.3

POS Tagging and Parsing

- Effect of Gold POS Tagging on LAS

WSJ22	*FootballDev*	*TwitterDev*
+ 1.1 - 2.0	+ 3.0 - 4.4	+ 7.9 - 11.3

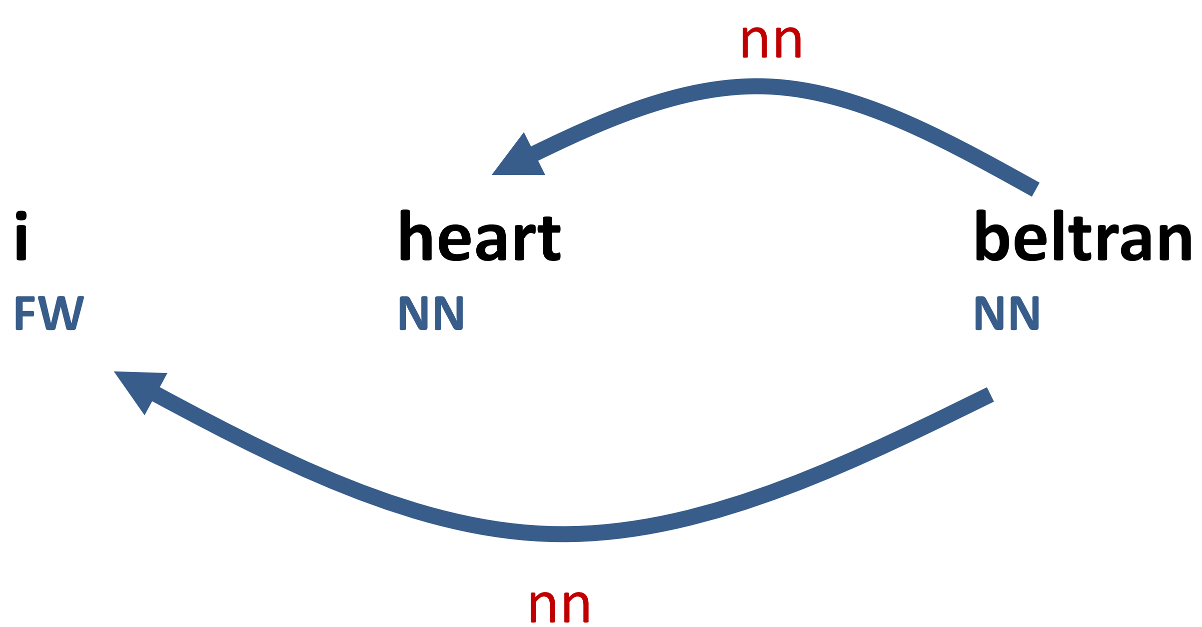
- LAS – UAS discrepancy

WSJ22	*FootballDev*	*TwitterDev*
~ 3	~ 4.5	~ 6

POS Confusion and Parsing

i	**heart**	**beltran**
FW	**NN**	**NN**

POS Confusion and Parsing



Making Use of Unlabelled Data

- Self-Training
  - Use trees parsed by a parser *P* to provide training material for *P* (McClosky et al. 2006, Huang and Harper 2009)

Making Use of Unlabelled Data

- Self-Training
  - Use trees parsed by a parser *P* to provide training material for *P* (McClosky et al. 2006, Huang and Harper 2009)
- Up-Training
  - Use a more accurate parser, *P1*, to provide training material for a less accurate parser, *P2* (Petrov et al. 2010)

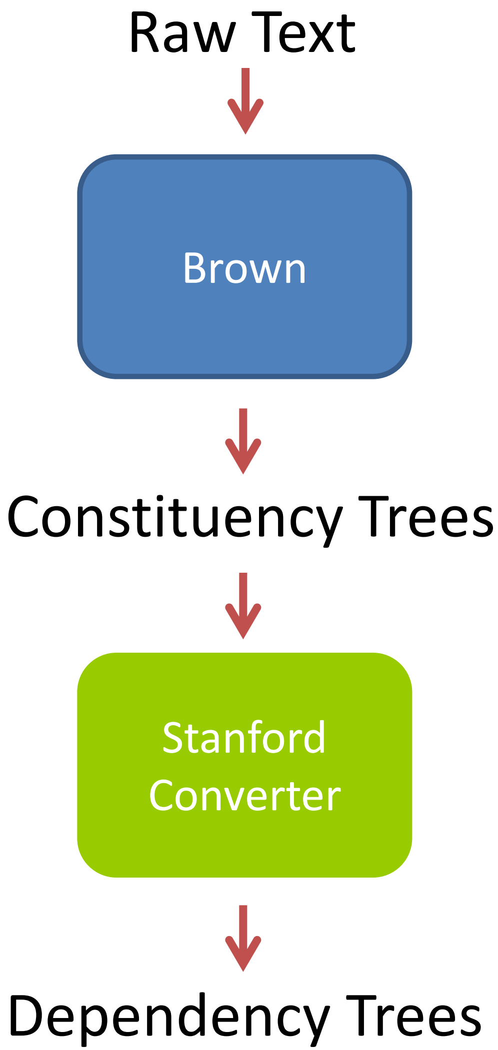
Making Use of Unlabelled Data

- Self-Training
  - Use trees parsed by a parser *P* to provide training material for *P* (McClosky et al. 2006, Huang and Harper 2009)
- Up-Training
  - Use a more accurate parser, *P1*, to provide training material for a less accurate parser, *P2* (Petrov et al. 2010)
  - Why not just use *P1*?

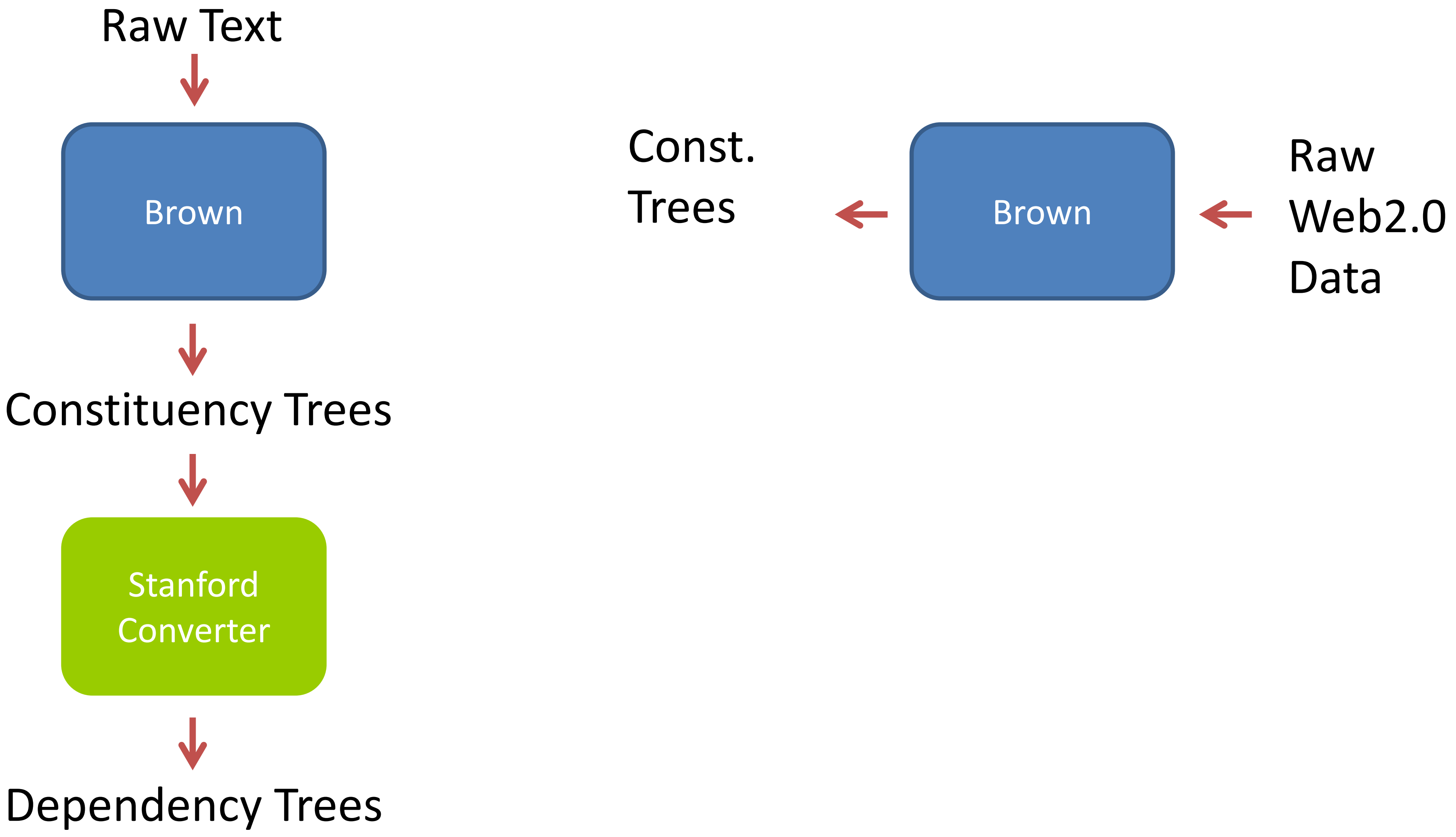
Making Use of Unlabelled Data

- Self-Training
  - Use trees parsed by a parser *P* to provide training material for *P* (McClosky et al. 2006, Huang and Harper 2009)
- Up-Training
  - Use a more accurate parser, *P1*, to provide training material for a less accurate parser, *P2* (Petrov et al. 2010)
  - Why not just use *P1*? *P2* is faster!

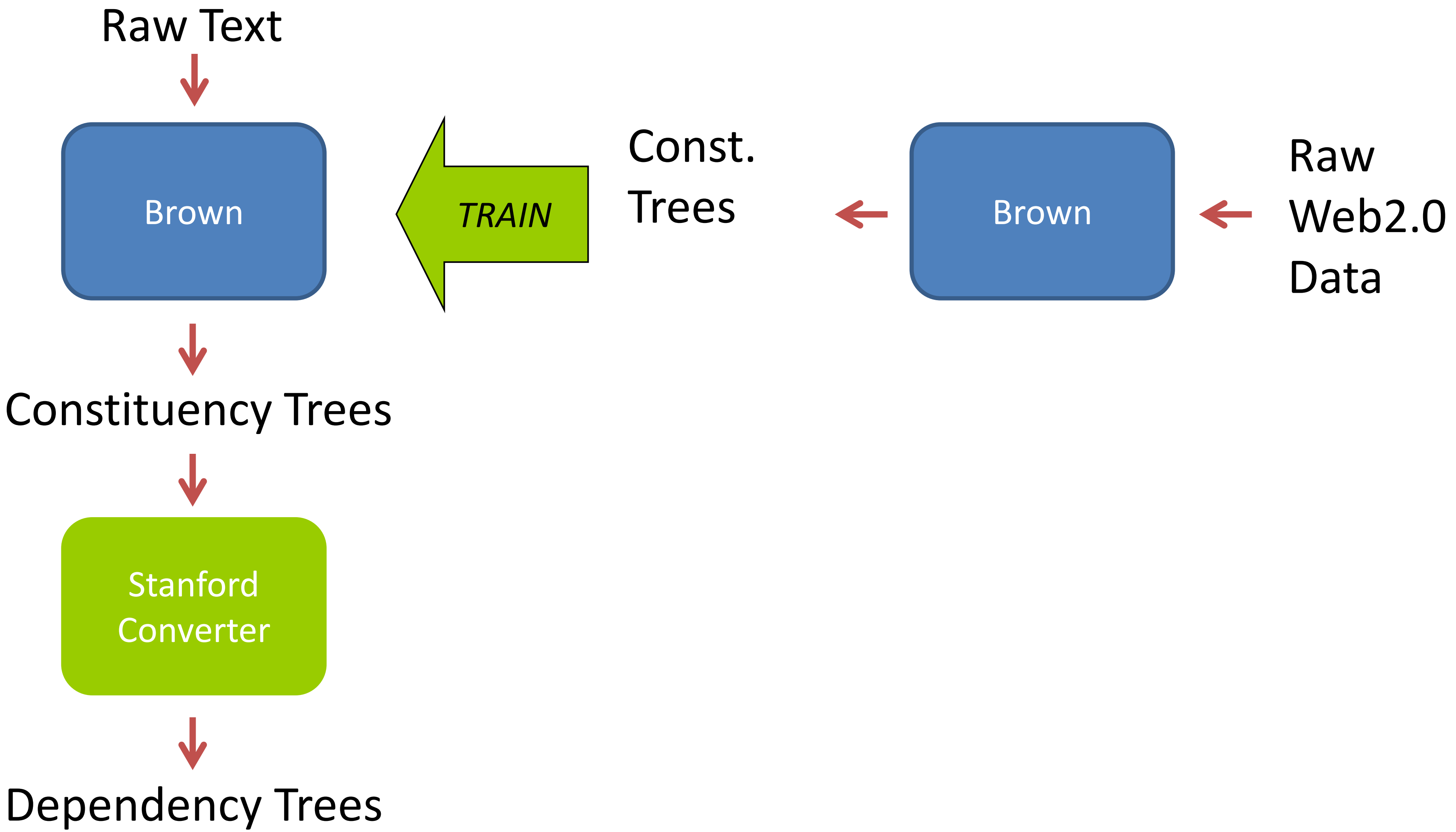
Self-Training



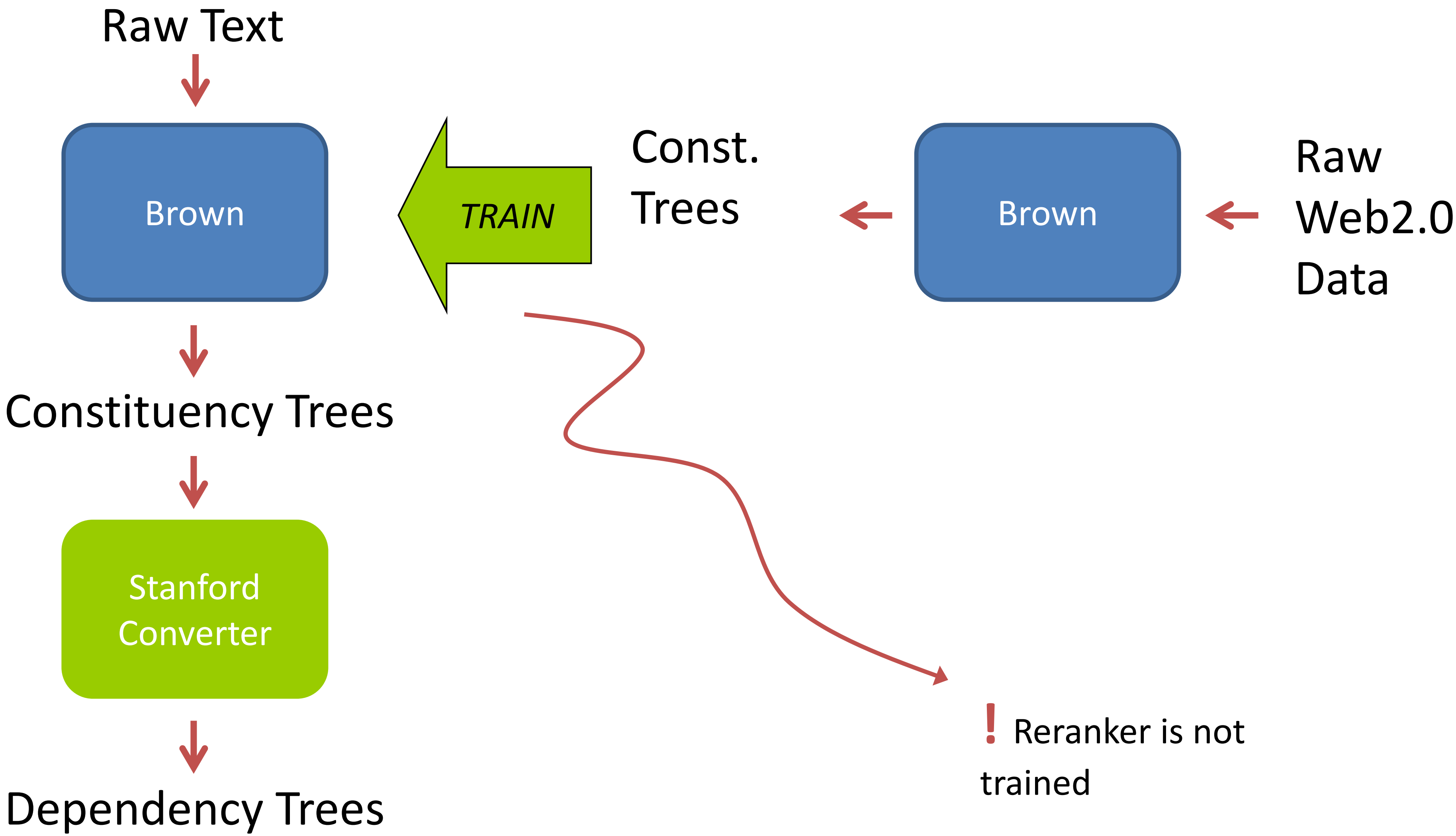
Self-Training



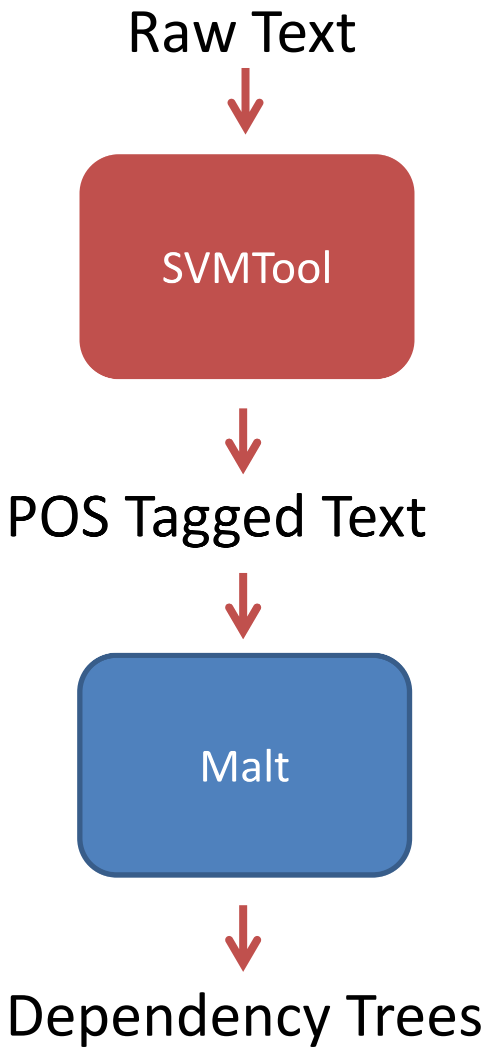
Self-Training



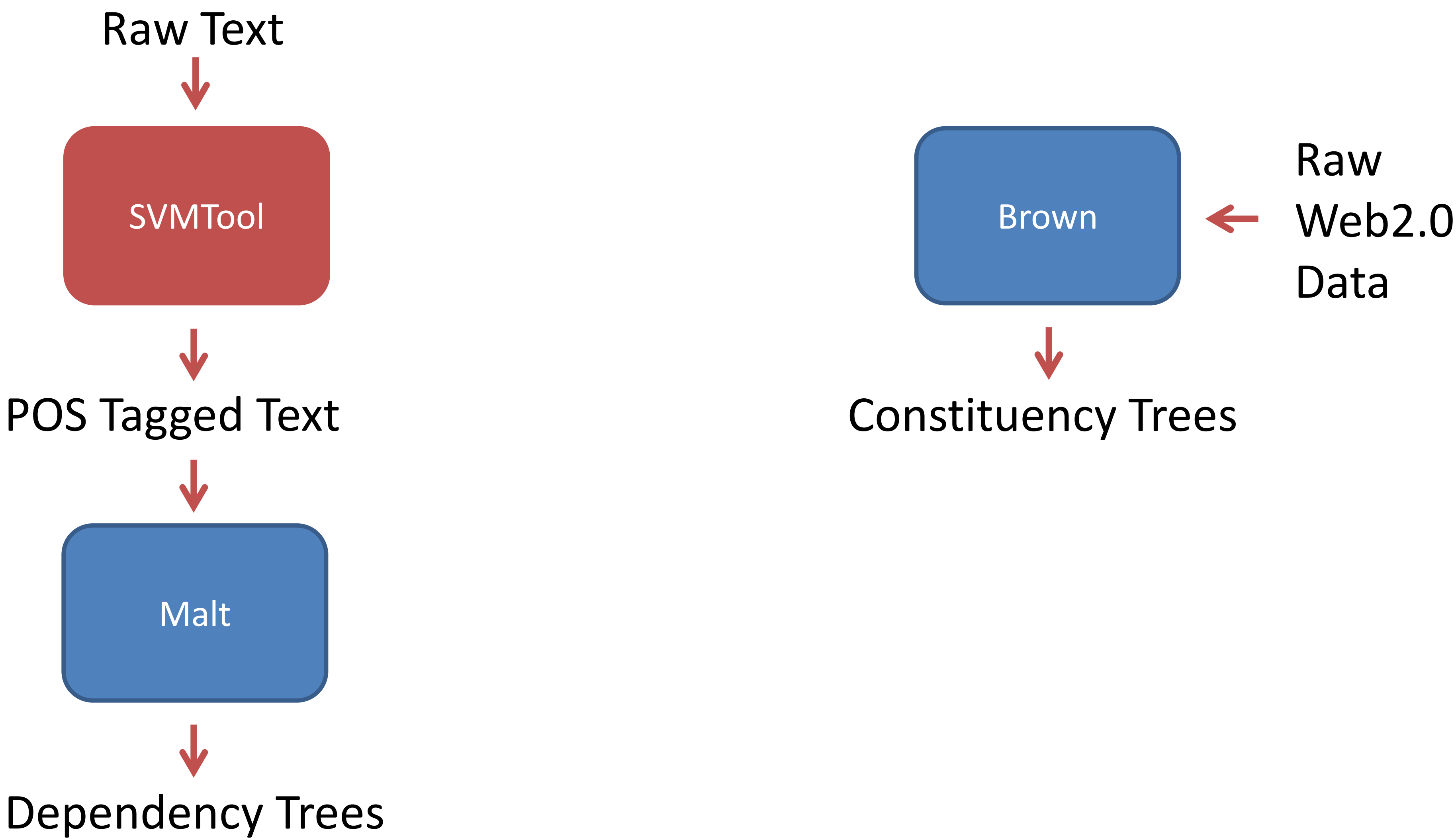
Self-Training



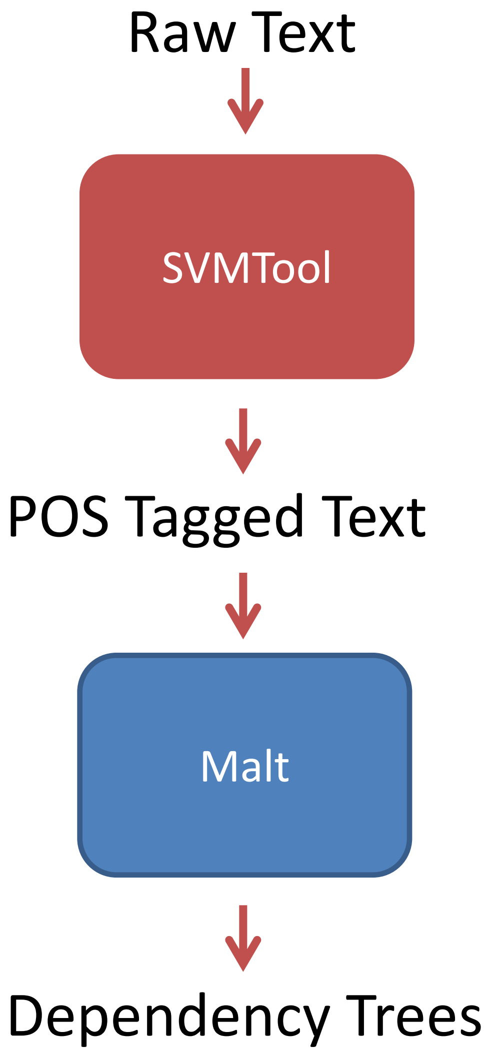
Vanilla Up-Training

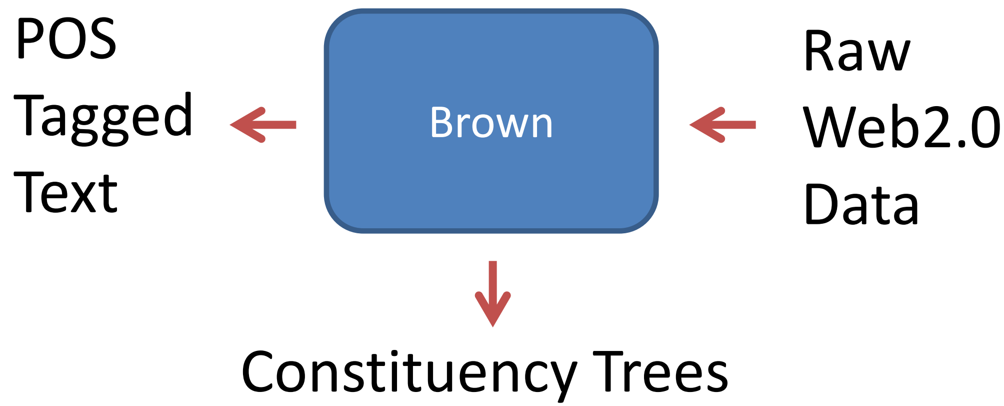


Vanilla Up-Training

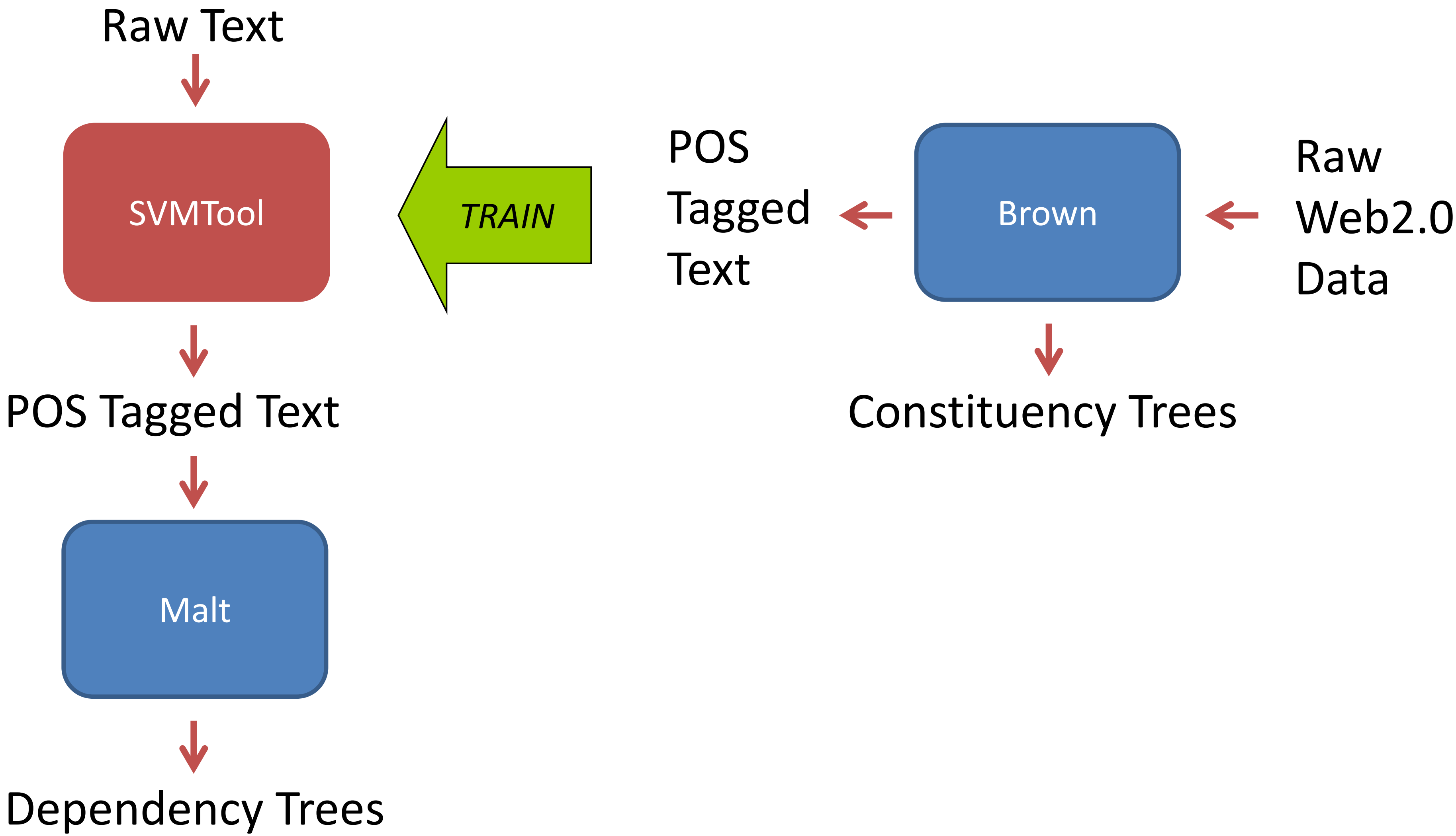


Vanilla Up-Training

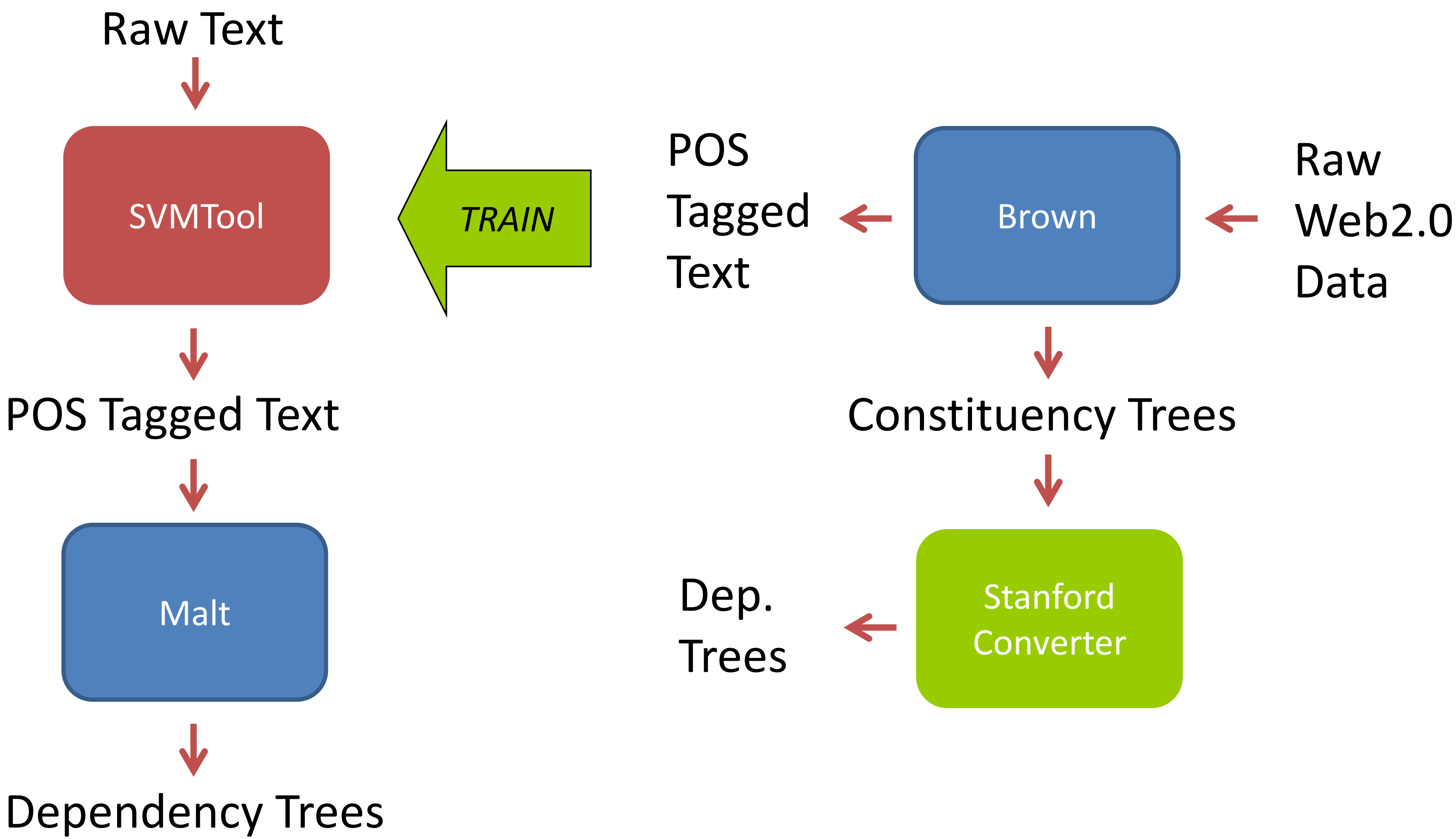




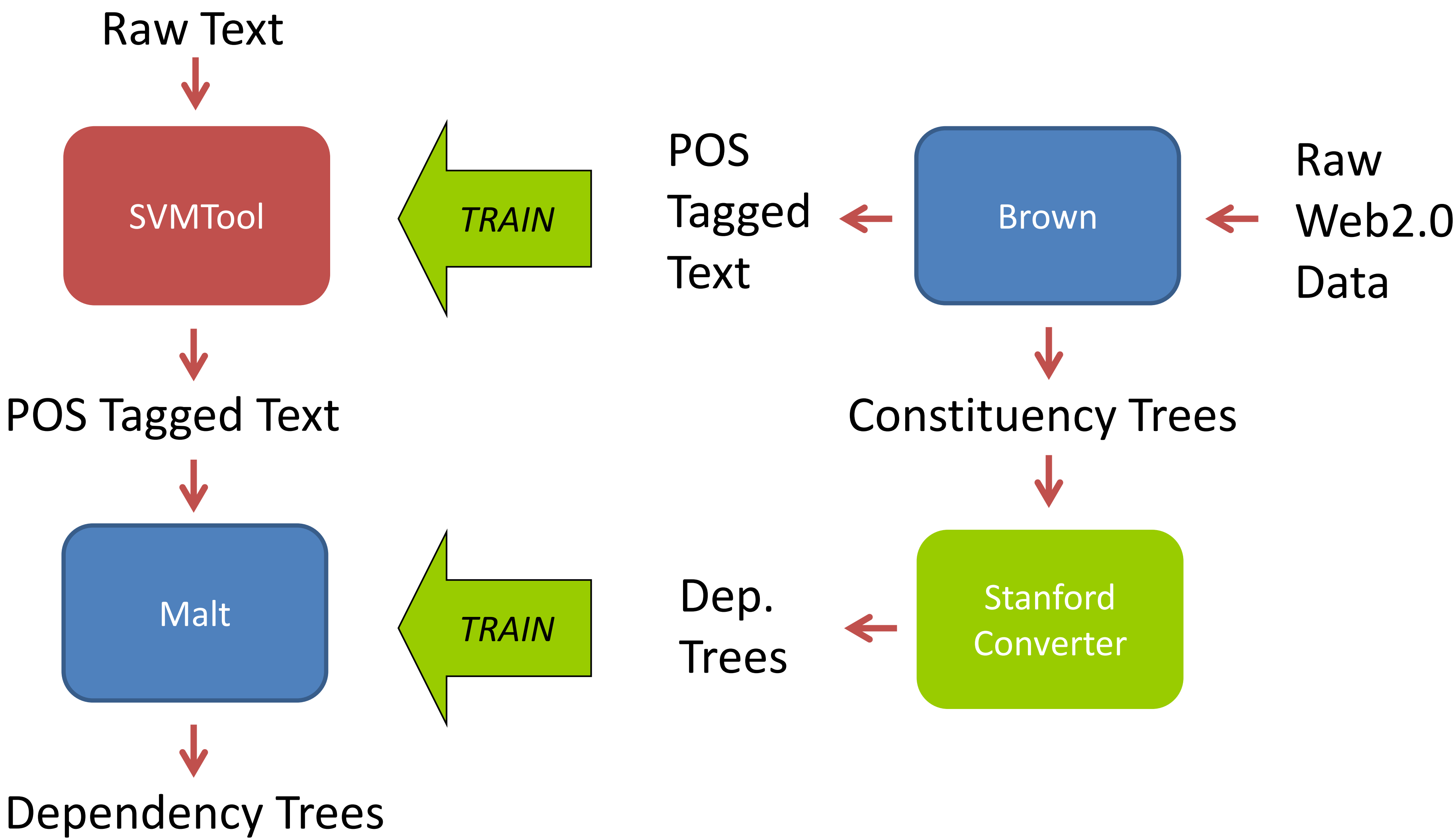
Vanilla Up-Training



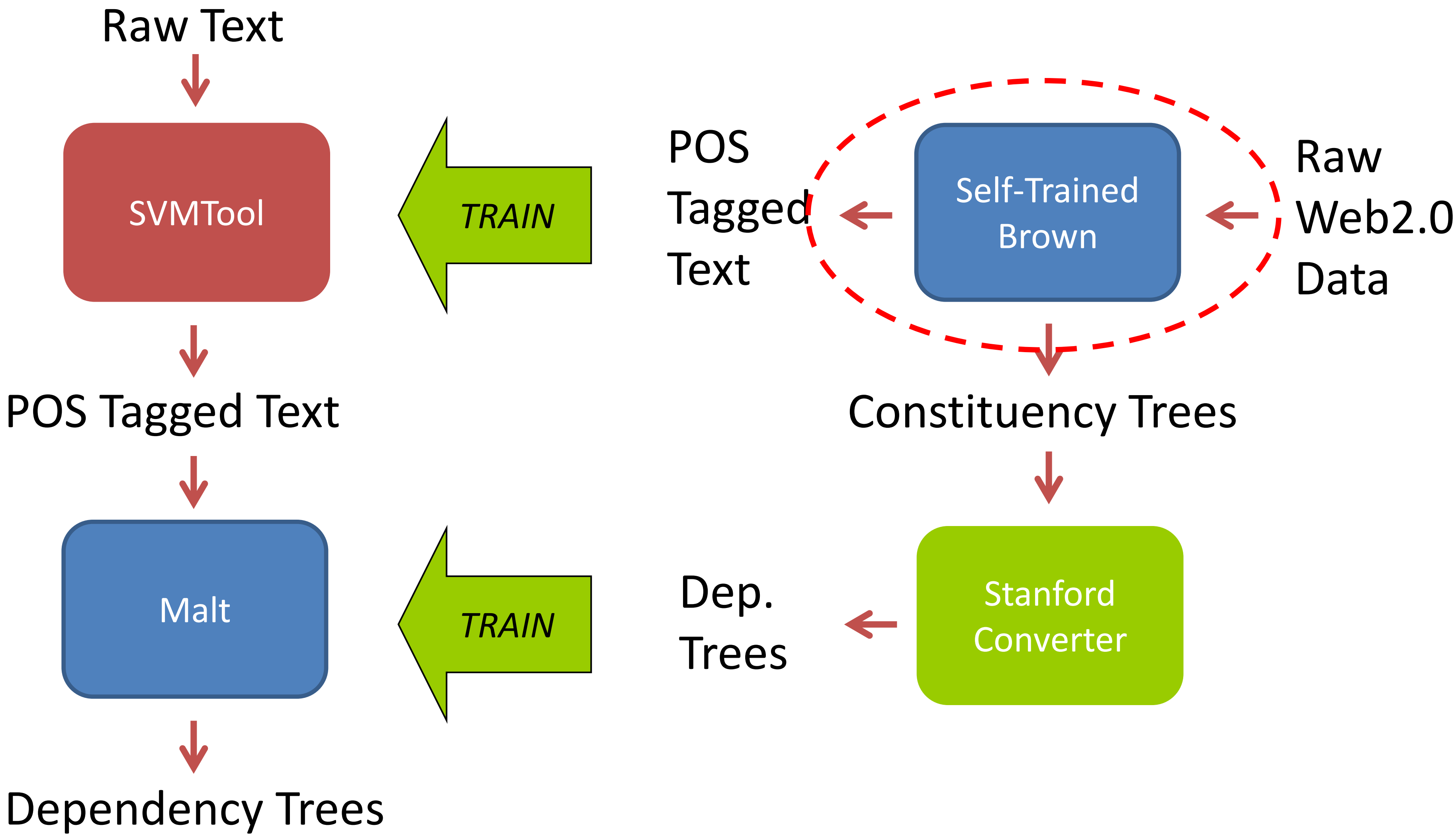
Vanilla Up-Training



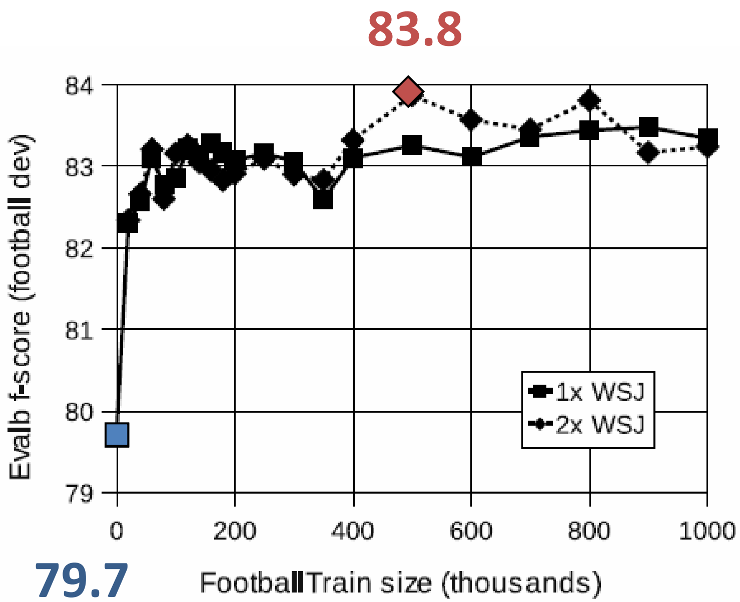
Vanilla Up-Training

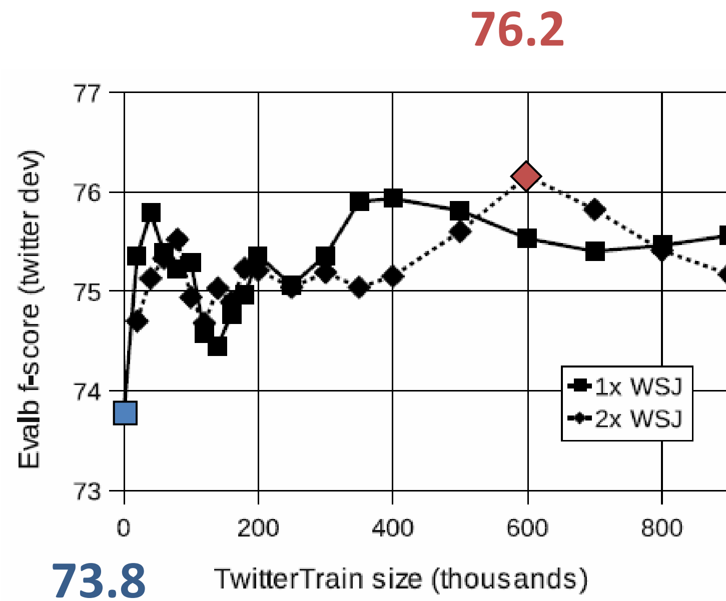


Domain Adapted Up-Training



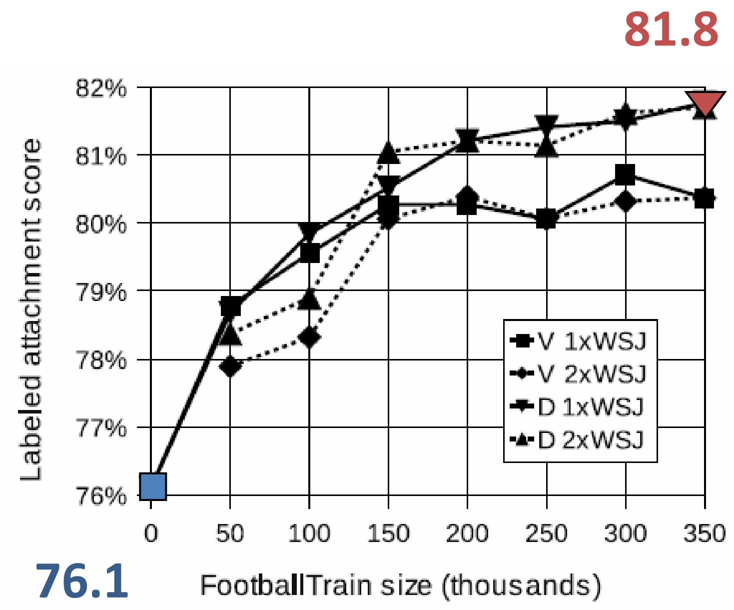
Self-Training Results

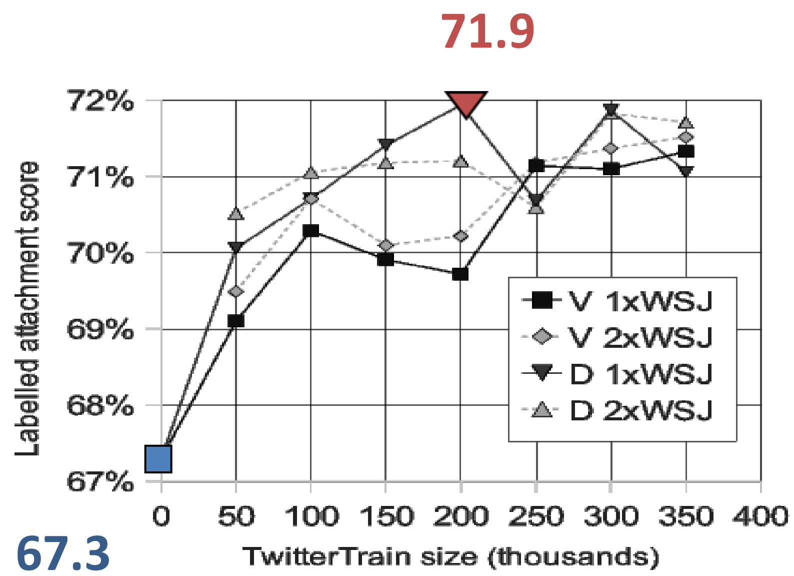




- Best *Football* grammar: 500K *FootballTrain* trees + 2 copies of WSJ2-21
- Best *Twitter* grammar: 600K *TwitterTrain* trees + 2 copies of WSJ2-21

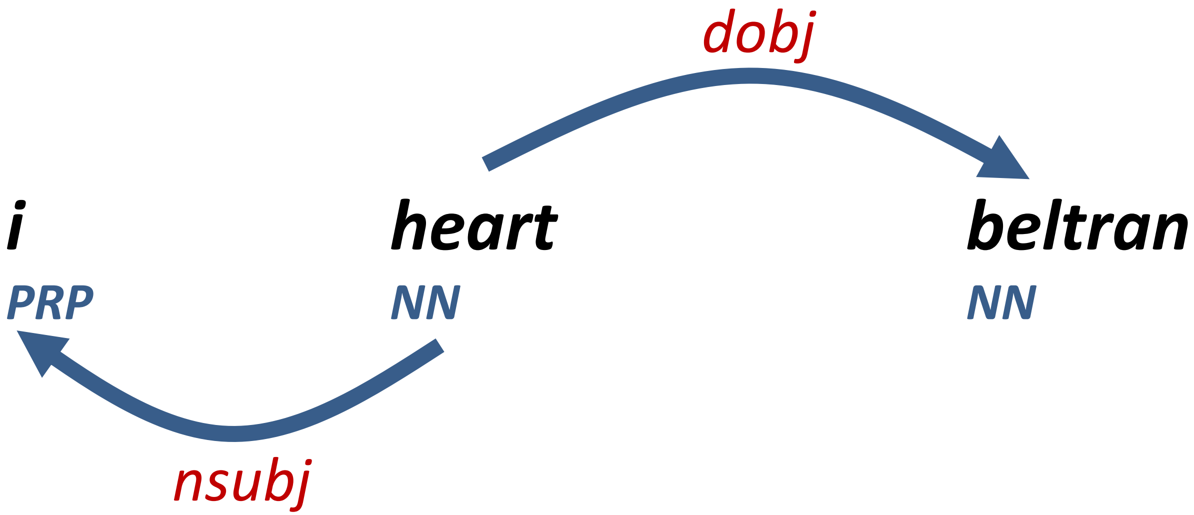
Up-Training Results





- Best *Football* grammar: 350K *FootballTrain* trees + 1 copy of WSJ2-21
- Best *Twitter* grammar: 200K *TwitterTrain* trees + 1 copy of WSJ2-21

Successful Retraining Example



Summary

- Introduced a new Web 2.0 dataset

Summary

- Introduced a new Web 2.0 dataset
- Detailed parser evaluation
  - 2.8 - 12.5 % drop in POS tagging accuracy
  - knock-on effect on parsing accuracy (9.5 - 21.7% drop)

Summary

- Introduced a new Web 2.0 dataset
- Detailed parser evaluation
  - 2.8 - 12.5 % drop in POS tagging accuracy
  - knock-on effect on parsing accuracy (9.5 - 21.7% drop)
- Investigated performance of existing unsupervised domain adaptation techniques

Summary

- Introduced a new Web 2.0 dataset
- Detailed parser evaluation
  - 2.8 - 12.5 % drop in POS tagging accuracy
  - knock-on effect on parsing accuracy (9.5 - 21.7% drop)
- Investigated performance of existing unsupervised domain adaptation techniques
- Introduced domain-adapted up-training

What next?

- Model combination (Petrov 2010, Surdeanu and Manning, 2010)

What next?

- Model combination (Petrov 2010, Surdeanu and Manning, 2010)
- Twitter-specific resources (Gimpel et al. 2011)

What next?

- Model combination (Petrov 2010, Surdeanu and Manning, 2010)
- Twitter-specific resources (Gimpel et al. 2011)
- Other parsers, other dependency schemes

What next?

- **Model combination (Petrov 2010, Surdeanu and Manning, 2010)**
- Twitter-specific resources (Gimpel et al. 2011)
- Other parsers, other dependency schemes

Part Three

Current Work

SANCL Shared Task

- Shared task on parsing the web
- Organised by Google
- New treebank
- 5 web genres (answers, blogs, emails, newsgroups, reviews)
- 2 sets of labelled data (blogs, emails) plus 5 sets of unlabelled data released in January for development
- 3 blind sets (answers, newsgroups, reviews) released one week before deadline

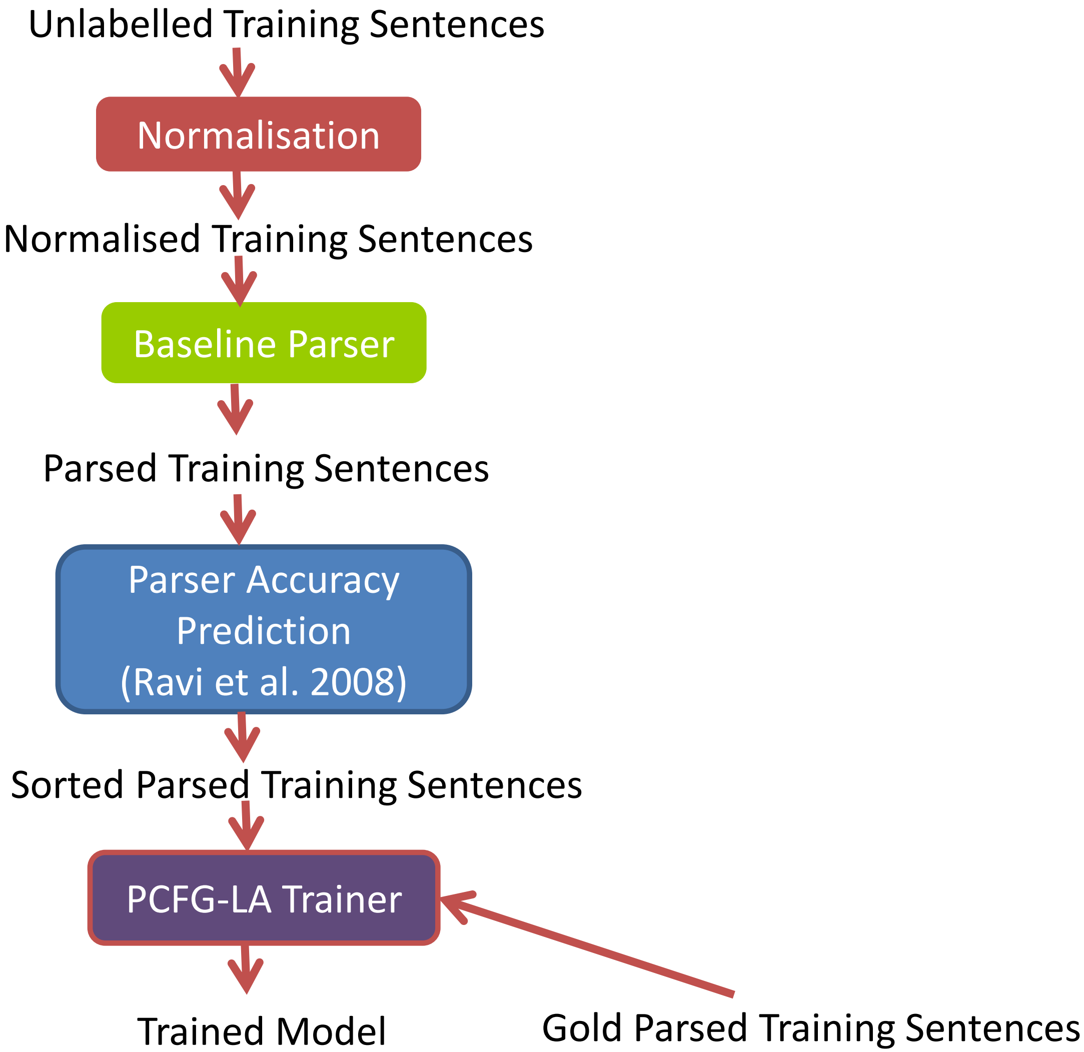
DCU-Paris 13 Team

1. Joseph Le Roux
2. Jennifer Foster
3. Joachim Wagner
4. Anton Bryl
5. Rasul Kaljahi

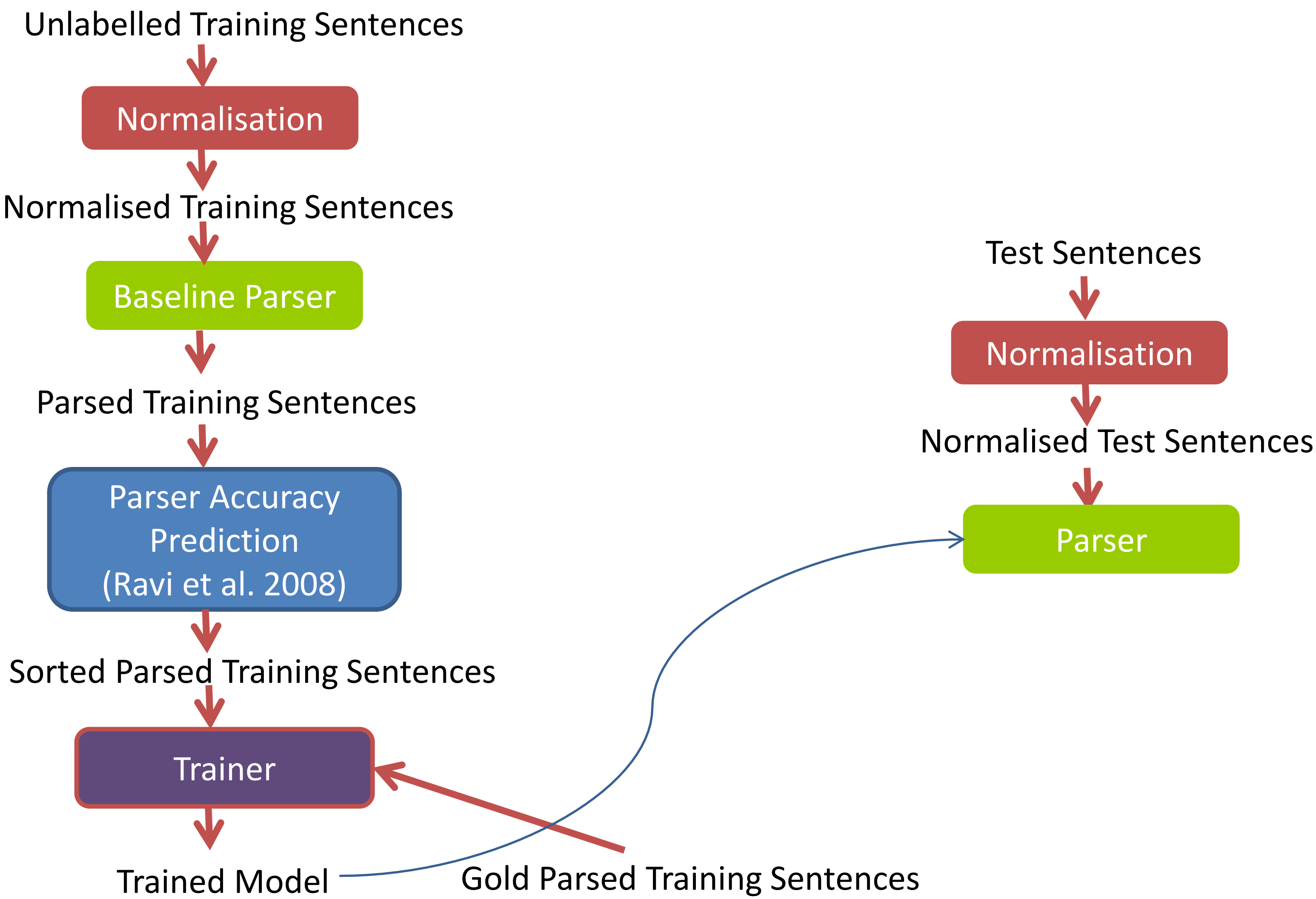
DCU-Paris 13 Systems

1. *LorgProdModel* (Constituent)
2. *CharniakCombination* (Constituent)
3. *CharniakCombinationVoting* (Dependency)

System Architecture



System Architecture



LorgProdModel

- Train 8 different PCFG-LA models (Petrov et al. 2006, Attia et al. 2010) on Ontonotes WSJ

LorgProdModel

- Train 8 different PCFG-LA models (Petrov et al. 2006, Attia et al. 2010) on Ontonotes WSJ
- Combine the grammars using a product model (Petrov 2010)

LorgProdModel

- Train 8 different PCFG-LA models (Petrov et al. 2006, Attia et al. 2010) on Ontonotes WSJ
- Combine the grammars using a product model (Petrov 2010)
- Parse the unlabelled data with the baseline product model grammar

LorgProdModel

- Train 8 different PCFG-LA models (Petrov et al. 2006, Attia et al. 2010) on Ontonotes WSJ
- Combine the grammars using a product model (Petrov 2010)
- Parse the unlabelled data with the baseline product model grammar
- Train 8 different self-trained models

LorgProdModel

- Train 8 different PCFG-LA models (Petrov et al. 2006, Attia et al. 2010) on Ontonotes WSJ
- Combine the grammars using a product model (Petrov 2010)
- Parse the unlabelled data with the baseline product model grammar
- Train 8 different self-trained models
- Combine the self-trained models using a product model (Huang et al. 2010)

LorgProdModel

- Train 8 different PCFG-LA models (Petrov et al. 2006, Attia et al. 2010) on Ontonotes WSJ
- Combine the grammars using a product model (Petrov 2010)
- Parse the unlabelled data with the baseline product model grammar
- Train 8 different self-trained models
- Combine the self-trained models using a product model (Huang et al. 2010)
- Computationally expensive - only 260k sentences from the unlabelled data could be used…

CharniakCombination

- Train several Brown first-stage models using the unlabelled data parsed using the *LorgProdModel* baseline grammar

CharniakCombination

- Train several Charniak first-stage models using the unlabelled data parsed using the LorgProdModel baseline grammar
- Training is quick – can use more data

CharniakCombination

- Train several Charniak first-stage models using the unlabelled data parsed using the LorgProdModel baseline grammar
- Training is quick – can use more data
- Combine the 50-best outputs of each grammar using a sentence-level product model

CharniakCombination

- Train several Charniak first-stage models using the unlabelled data parsed using the LorgProdModel baseline grammar
- Training is quick – can use more data
- Combine the 50-best outputs of each grammar using a sentence-level product model
- For each sentence, multiply the parse probabilities for the trees produced for that sentence by each of the models

CharniakCombination

- Train several Charniak first-stage models using the unlabelled data parsed using the LorgProdModel baseline grammar
- Training is quick – can use more data
- Combine the 50-best outputs of each grammar using a sentence-level product model
- For each sentence, multiply the parse probabilities for the trees produced for that sentence by each of the models
- Output the tree with the highest probability

CharniakCombinationVoting

- Take the trees produced by three different Brown combined systems

CharniakCombinationVoting

- Take the trees produced by three different Brown combined systems
- Convert them to dependencies (Stanford converter)

CharniakCombinationVoting

- Take the trees produced by three different Brown combined systems
- Convert them to dependencies (Stanford converter)
- Combine the dependency trees using a simple voting algorithm (Surdeanu and Manning, 2010)

Full Set of Results

SYSTEM	Answers		Newsgroups		Reviews		WSJ		Average Web	
Baseline	*75.92*	90.20	*78.14*	*91.24*	*77.16*	*89.33*	*88.21*	*97.08*	*77.07*	*90.26*
LorgProdModel	82.19	91.63	84.33	93.39	84.03	92.89	90.53	97.53	83.52	92.64

https://sites.google.com/site/sancl2012/home/shared-task/results

ConfidentMT Project

- Improve the accuracy of machine translated Symantec customer forum data

ConfidentMT Project

- Improve the accuracy of machine translated Symantec customer forum data
- Customers are bypassing traditional help services and helping each other via customer forums

ConfidentMT Project

- Improve the accuracy of machine translated Symantec customer forum data
- Customers are bypassing traditional help services and helping each other via customer forums
- English forum data is plentiful

ConfidentMT Project

- Improve the accuracy of machine translated Symantec customer forum data
- Customers are bypassing traditional help services and helping each other via customer forums
- English forum data is plentiful
- Could this English data be useful to Symantec's French and German customers?

Confident MT Project

Can we use domain-adapted parsers to build better syntax-augmented SMT systems?

Confident MT Project

Can we use domain-adapted parsers to build better syntax-augmented SMT systems?

To be continued….

Thanks!
Questions?