Word Sketches from Other Parsers: CONLL format in Sketch Engine

Siva Reddy Diana McCarthy

Lexical Computing Ltd., UK

2nd International Sketch Engine Workshop
March 17 2011

Sketch Engine

Outline

1. Background and Motivation
2. Existing parsers
3. Demo

Sketch Engine

Background and Motivation

Background: Sketch Grammar

- Current word sketches are built using Sketch Grammar
- Manually written by ourselves, collaborators and language experts
- Lot of work on us. Expensive!!

Sketch Engine

Background and Motivation

Background: Sketch Grammar

- Current word sketches are built using Sketch Grammar
- Manually written by ourselves, collaborators and language experts
- Lot of work on us. Expensive!!

Motivation: Benefit from the research in NLP

- We have a strong team in NLP
- Many research groups have their own dependency parsers
- Lots of research already done. Time to harvest and eat the fruits
- Scale to as many languages as possible

Sketch Engine

Existing Parsers

Shared tasks

- Competitions for building new and efficient parsers.
- Around 20 languages
- CONLL Shared task: Arabic, Bulgarian, Chinese, Czech, Danish, Dutch, German, Japanese, Portuguese, Slovene, Swedish, Spanish, Turkish
- ICON Shared task: Hindi, Telugu, Bengali
- Training data is available for download
- Malt and MST Parser work for many languages

Sketch Engine

Existing Parsers

Shared tasks

- Competitions for building new and efficient parsers.
- Around 20 languages
- CONLL Shared task: Arabic, Bulgarian, Chinese, Czech, Danish, Dutch, German, Japanese, Portuguese, Slovene, Swedish, Spanish, Turkish
- ICON Shared task: Hindi, Telugu, Bengali
- Training data is available for download
- Malt and MST Parser work for many languages

Existing Parsers

- Malt, MST, RASP, Stanford, Collins, Constraint-based . . .
- Most parsers support CONLL format

Sketch Engine

CONLL format

Very similar to Sketch Engine vertical format

```
1   Ze              ze          Pron  per|3|evofmv|nom          2   su
2   had             heb         V     trans|ovt|1of2of3|ev      0   ROOT
3   met             met         Prep  voor                      8   mod
4   haar            haar        Pron  bez|3|ev|neut|attr        5   det
5   moeder          moeder      N     soort|ev|neut             3   obj1
6   kunnen          kan         V     hulp|ott|1of2of3|mv       2   vc
7   gaan            ga          V     hulp|inf                  6   vc
8   winkelen        winkel      V     intrans|inf               11  cnj
9   ,               ,           Punc  komma                     8   punct
10  zwemmen         zwem        V     intrans|inf               11  cnj
11  of              of          Conj  neven                     7   vc
12  terrassen       terras      N     soort|mv|neut             11  cnj
13  .               .           Punc  punt                      12  punct
```

Sketch Engine

Word Sketches compiled from CONLL format

UKWaC parsed with Malt Parser

- Word Sketch from CONLL:
  `http://the.sketchengine.co.uk/auth/preloaded_corpus/pukwac/ske/wsketch_form`
- Word Sketch from Sketch Grammar:
  `http://the.sketchengine.co.uk/auth/preloaded_corpus/ukwac2/ske/wsketch_form`
- High similarity is observed.

Sketch Engine

Word Sketch of the verb **play**

CONLL

OBJ	17808	5.2
role	3015	10.0
game	1285	8.83
football	305	8.56
part	1675	8.15
host	230	8.02
guitar	189	7.96
piano	114	7.49
instrument	154	7.34
music	289	7.29
cricket	94	7.26
gig	106	7.17
ball	116	6.79
tune	81	6.78
golf	68	6.72
rugby	60	6.67
tennis	61	6.66
match	109	6.64
poker	77	6.63
song	140	6.6
drum	56	6.49
squash	47	6.36
sport	91	6.35
bass	50	6.25
card	140	6.16
chess	37	6.04

SkE Grammar

object	253889	4.2
role	47209	10.9
game	20670	9.55
part	28097	9.31
football	5522	8.95
guitar	3740	8.5
host	3367	8.3
field	5448	8.0
music	5836	7.94
match	2878	7.86
piano	1941	7.79
gig	2013	7.74
instrument	2444	7.68
cricket	1500	7.44
golf	1587	7.43
ball	2070	7.34
tune	1543	7.34
song	2534	7.3
rugby	1259	7.22
drum	1152	7.08
card	2775	7.07
bass	1068	6.9
poker	1799	6.87
tennis	999	6.85
sport	1547	6.75
trick	828	6.52

Sketch Engine

Word Sketch of the verb **play**

CONLL

SBJ	11527	2.4
who	678	7.1
he	709	6.67
band	106	6.26
they	578	6.24
she	156	5.99
we	607	5.89
myself	35	5.88
him	90	5.56
I	440	5.11
actor	21	5.09
Shakespeare	14	5.03
kid	21	4.93
them	102	4.85
you	441	4.81
ombudsman	10	4.78
Band	11	4.77
player	45	4.63
childrens	9	4.62
Club	14	4.54
Mystery	8	4.47
orchestra	9	4.41
that	194	4.39
Mark	11	4.34
i	21	4.33
team	69	4.31

Sketch Grammar

subject	89717	2.6
band	3276	7.98
actor	1345	7.97
musician	716	7.19
actress	370	6.82
player	1653	6.73
kid	513	6.4
guy	463	6.32
team	2206	6.26
guitar	407	6.25
boy	654	6.22
orchestra	248	6.17
level	1901	5.79
childrens	158	5.77
music	951	5.66
game	1049	5.54
girl	415	5.51
lad	170	5.49
child	2094	5.47
piano	148	5.31
guitarist	125	5.3
fun	299	5.21
everyone	382	5.16
youngster	124	5.08
radio	227	5.04
ensemble	107	5.03

Sketch Engine

Thank you

Sketch Engine