Universal Sketch Grammar

Siva Reddy, Adam Kilgarriff, Pavel Rychlý

Lexical Computing Ltd.

Skew-3 Workshop, Brno, Czech Republic
Lexical Computing Ltd.
March 21 2012

Sketch Engine

Current State of Corpora in SkE

Total Corpora	200
Languages	63
Languages with POS Taggers	25
Languages with Sketch Grammar	24

Sketch Engine

Current State of Corpora in SkE

Total Corpora	200
Languages	63
Languages with POS Taggers	25
Languages with Sketch Grammar	24

Long Term Ambition

- Unified tagging architecture for all languages
- Word Sketches for every language

Sketch Engine

Current State of Corpora in SkE

Total Corpora	200
Languages	63
Languages with POS Taggers	25
Languages with Sketch Grammar	24

Long Term Ambition

- Unified tagging architecture for all languages
- Word Sketches for every language

Our Current Solution

- Universal Sketch Grammar
- Works with any corpora

Sketch Engine

Universal Sketch Grammar

Corpus with no POS tags

- Tag 200 most freq words in language as FREQ
- Numerals as CRD
- Punctuations as PUNC
- Other words as CONTENT
- Sketch Grammar capturing word-associations

Sketch Engine

Universal Sketch Grammar

Corpus with no POS tags

- Tag 200 most freq words in language as FREQ
- Numerals as CRD
- Punctuations as PUNC
- Other words as CONTENT
- Sketch Grammar capturing word-associations

- Grammar: `http://corpdev.sketchengine.co.uk/run.cgi/wsdef?corpname=c257ea9d`
- Corpus: `http://corpdev.sketchengine.co.uk/run.cgi/first_form?corpname=c257ea9d`

Sketch Engine

Tagged Universal Sketch Grammar

Corpus with POS tags but not Sketch Grammar

- Sketch Grammar capturing type based word-associations
- Currently for Malay, Indonesian and other languages to come.

Sketch Engine

Tagged Universal Sketch Grammar

Corpus with POS tags but not Sketch Grammar

- Sketch Grammar capturing type based word-associations
- Currently for Malay, Indonesian and other languages to come.

- Grammar: `http://corpdev.sketchengine.co.uk/run.cgi/wsdef?corpname=6cbce6b8`
- Corpus: `http://corpdev.sketchengine.co.uk/run.cgi/first_form?corpname=6cbce6b8`

Sketch Engine

Universal Word Sketches in Action

- Indonesian and Malay Lexicography
- Our study on Turkish WordNet topic coherence
  - External task based evaluation for word sketches
  - Thesaurus almost as accurate as from sketch grammar??

Sketch Engine

Comparison: Thesaurus of house in different corpora

ukWaC freq = 392287

Lemma	Score	Freq
building	0.534	363768
home	0.483	675005
room	0.461	364176
garden	0.44	171248
church	0.432	253000
shop	0.421	171029
town	0.413	260679
property	0.412	329119
area	0.409	1103121
office	0.407	289728
village	0.398	169340

UKWaC Tagged Universa

Lemma	Score	Freq
building	0.355	4925
home	0.318	8319
place	0.31	12239
room	0.303	4670
day	0.293	18961
thing	0.278	12738
country	0.277	11800
time	0.277	36561
site	0.276	12919
area	0.276	16055
town	0.275	3473

UKWaC Untagged Universa

Lemma	Score	Freq
building	0.386	4925
home	0.345	8760
place	0.339	15708
time	0.327	36742
day	0.325	18961
year	0.324	33366
up	0.319	33776
room	0.317	4673
well	0.316	26994
one	0.316	45910
area	0.314	16055

Sketch Engine