Finding commonest match for word sketch collocations

Vít Baisa

Lexical Computing

SKEW3, 2012, Brno

Introduction

- intended as helpful feature for WS evaluators:
  see `has_obj` **final**

Introduction

- intended as helpful feature for WS evaluators:
  see `has_obj` **final** $\rightarrow$ **saw** *world cup* **finals**

Introduction

- intended as helpful feature for WS evaluators:
  see `has_obj` **final** $\rightarrow$ **saw** *world cup* **finals**

- discover a common string for collocates
- using left and right context
- include intervening words

Algorithm

Input: a pair of lemmas from a word sketch

Init: from corresponding concordance lines, use strings from the left lemma to the right lemma (*match*) plus 3 left and 3 right words

```
if the most freq. match occurs > N/4 times:
    commonest_match = the most frequent match
    n = freq(commonest_match)
    if the most freq. L/R extension occurs > n/4 times:
        extended_commonest_match = the most freq. L/R extension
        n = freq(extended_commonest_match)
        if the most freq. L/R extension occurs < n/4:
            return extended_commonest_match
        else:
            extend recursively L/R, up to 3 words
    else:
        return commonest_match
else:
    return ""
```

Example I

Word sketch for lemma **morning**

modifier	tomorrow	1422	tomorrow morning [93.0 %]
modifier	next	8722	next morning [87.5 %]
modifier	yesterday	1267	yesterday morning [84.1 %]
modifier	coffee	1254	coffee morning [49.2 %]
modifier	early	4387	early morning [77.3 %]
modifier	sunny	490	sunny morning [85.1 %]
modifier	weekday	289	weekday morning [62.3 %]
modifier	mid	297	mid morning [86.9 %]

Example II

Extended commonest matches for collocates from various gramrels.

lemma	collocate	commonest	%
evaluation	histological	eosin for histological evaluation	37.5
evaluation	determinant	evaluation of determinant	40.0
key	backspace	TAB and BACKSPACE keys can be used	30.0
key	keyboard	key on the keyboard	27.3
squirrel	golden-mantled	golden-mantled ground squirrels	33.3
squirrel	popper	squirrels in a popcorn popper	27.3
republic	USSR	republics of the USSR	43.4
bride	toast	toast to the bride and groom	55.6
marry	subsequently	whom he subsequently married	40.0

What it is good for?

- provide evaluators with explanatory examples
- provide the most common utterance of the collocates
- Good Dictionary EXample for the collocates
- discover obvious multiword expressions

Conclusion, future plans

- accessible on `beta.sketchengine.co.uk` (**Sketch-Eval**)
- computed for top N collocates, on the fly

- make it faster
- pre-compute commonest matches
- show within WS by default
- get feedback from users
- tune parameters, or provide a way to change them