

# SENSE CHANGES & MULTI-WORD EXPRESSIONS



Martin Emms and Arun Jayapal

Dept of Computer Science, Trinity College, Dublin, Ireland  
mtemms@scss.tcd.ie, jayapala@scss.tcd.ie

## Introduction

A n-gram might be termed a multiword expression (MWE) because it has a property, possibly a meaning, which is odd given its parts.

a *further twist* is that in one context a *token* of the n-gram *type* might exhibit the irregular MWE usage, but in a different context it might not.

Still a *further twist* is added by language *change*: quite possibly the irregular MWE usage *emerged*, and was predated by only more transparent usages.

Following on from [1], we propose a method to detect such cases of emergence by automatic means.

## Motivation & Use case

1995 *the wind lifted his three-car garage and smashed it to the ground.*  
S=1: 'destruction' i.e. transparent

2013 *sensational group CEO, totally smashed it in the BGT (Britain Got Talent)*  
S=2: 'excelled' i.e. opaque

The (S=2, 'excelled') usage of *smashed it* is a *recent possibility*, and was *predated* by an era in which it was not possible.

We would like to detect this by **unsupervised means from time-stamped text**

Other examples

1990: *give me time vs*  
2013: *enjoy some me time*  
1995: *going forward from the entrance vs*  
2009: *going forward, the company should*

Possibly useful for detecting that aligned data for an SMT system is out of date and may mis-translate recent examples.

Google translations into German  
"and smashed it to the ground." (from 1995, standard destructive usage  
→ "und schlug ihn zu Boden" — correct translation  
"the sensational group totally smashed it!" (from 2013, meaning 'excelled')  
→ "die sensationelle Gruppe völlig zertrümmert es!" — poor translation

## References

[1] Martin Emms. Dynamic EM in Neologism Evolution. In *Proceedings of IDEAL*, 2013.

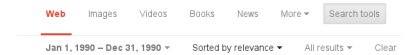
## Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

## Data and Algorithms

1995 *the wind lifted his three-car garage and smashed it to the ground.*  
:  
1996  
:  
2013 *sensational group CEO, totally smashed it in the BGT (Britain Got Talent)*  
:  
:

data via Google's *custom date range*



for given n-gram download 100 hits for **year-long** time-spans from 1990–2013 (total ≈ 2k hits per n-gram)

treat each n-gram occurrence as  $\langle Y, S, \vec{W} \rangle$  where

$\vec{W}$  = words in the context,  $Y$  = year of occurrence,  $S$  = the usage variant  
— in the downloaded data  $S$  is **hidden**.

Define model of  $p(Y, S, \vec{W})$  via (2) below

$$\begin{aligned} p(Y, S, \vec{W}) &= p(Y)p(S|Y)p(\vec{W}|S, Y) \\ &\approx p(Y)p(S|Y)p(\vec{W}|S) \quad (1) \\ &\approx p(Y)p(S|Y)\prod_i p(W_i|S) \quad (2) \end{aligned}$$

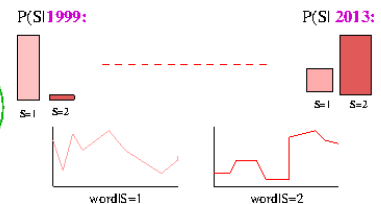
**Key temporal feature:** for the usage-variant  $S$ , instead of a single prior  $p(S)$ ,  $p(S|Y)$  is effectively a *succession of priors*

**Conditional words/time independence** line (1) assumes that **given S**, context words ( $\vec{W}$ ) and time  $T$  are independent.

Model parameters usage-given-time  $p(S|Y)$  and word-given-usage  $p(W_i|S)$  are estimated with an Expectation-Maximisation technique

time-stamped raw text

1999: ..... smashed it ....  
:  
:  
:  
2013 ..... smashed it ....



**E-step** populates a table  $\gamma$ , such that for each data point  $d$ , and possible  $S$  value  $s$ ,  $\gamma[d][s]$  stores  $P(S = s|Y = y^d, \vec{W} = \vec{w}^d)$ .

$$P(S = s|Y = y) = \frac{\sum_d (\text{if } Y^d = y \text{ then } \gamma[d][s] \text{ else } 0)}{\sum_d (\text{if } Y^d = y \text{ then } 1 \text{ else } 0)} \quad (1)$$

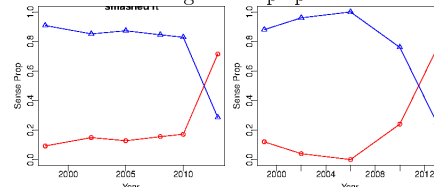
**M-step** based on  $\gamma$ , fresh parameter values are re-estimated according to update formulae 1 and 2

$$P(w|S = s) = \frac{\sum_d (\gamma[d][s] \times \text{freq}(w \in \vec{W}^d))}{\sum_d (\gamma[d][s] \times \text{length}(\vec{W}^d))} \quad (2)$$

## Outcomes

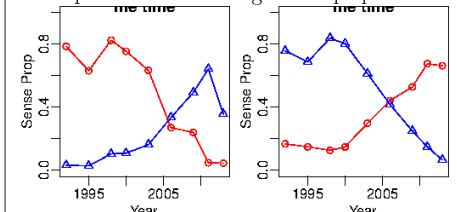
*smashed it* (empirical and unsupervised)

○ = 'excelled' usage in emp. plot



*me time* (empirical and unsupervised)

△ is 'personal time' usage in emp. plot



**Figure 1:** outcomes for  $p(S|Y)$  for *smashed it* and *me time*. **unsupervised** = EM estimate. **empirical** = an ML estimate based on a 10% hand-annotated subset.

In empirical plots, for *smashed it* and *me time* the MWE-usages have an upward trend, as expected. Given that with the unsupervised method it is indeterminate which value of  $S$  may correspond to a given objectively real usage, the unsupervised plots broadly concur with the empirical.

Unsup favoured vocab apparent recent usage of *smashed it*: *!!*, *guys*, *really*, *completely*, *They*, *!*  
Unsup favoured vocab apparent old usage of *smashed it*: *smithereens*, *bits*, *bottle*, *onto*, *phone*

We have looked also at the expressions *biological clock* and *going forward*, finding similar empirical and observed emergence of a recent novel usage. Ongoing we would like to consider possibly related other methods and different time stamped corpora.