Sense changes and Multiword Expressions

Martin Emms and Arun Jayapal
Dept of Computer Science, Trinity College, Ireland

February 25, 2014

Sometimes a multiword expression (MWE) will be perceived as such because it seems to have a meaning which is not entirely, or even at all, predictable based on its component words. Also, like single-word expressions, the word-sequence of a given MWE may not in all contexts contribute to the overall meaning in the same way: in some contexts its contribution is compositional via possible senses of the individual words – in short it is not really a MWE in those contexts – and in other contexts the contribution is via the special multiword meaning. The examples below, illustrate this.

(a) *and* **smashed it** *to the ground.*
(a′) *... sensational group CEO, totally* **smashed it** *in the BGT*
(b) *my schedule gave* **me time** *to get adjusted*
(b′) *it's important to set time out and enjoy some* **me time**

In (a) the semantic contribution of the n-gram *smashed it* is straightforwardly via its parts, whilst in (a′), it is not, with it taken as roughly synonymous with 'excelled', *smashed* not contributing its standard destructive sense, nor *it* referring to some previously described object. In (b) the parts of the n-gram *me time* are separate, simple, dependants of *gave*, but in (b′), it is a noun-phrase, meaning something like 'personal time'; a marker of its irregularity is that addressed to someone the *me* in (b′) could refer to the addressee. Language is a dynamic phenomenon, and mwe-usages shown above (primed) are illustrations of relatively recent *innovations* in English. Following on from [Emms, 2013], the work reported here is concerned with whether automatic, unsupervised means can be found by which it is possible to detect the emergence of these kinds of new, mwe-usages.

To do this a time-stamped corpus is necessary. The approach we have taken to this is to exploit a facility that Google has offered for some time – *custom date range* – whereby it is possible to specify a time period for searched documents. For a given n-gram we repeatedly set different year-long time spans (from 1993 to 2013), saving the first 100 returned hits as examples of the expression's use. Subsequently these were merged to give spans of 3 years duration.

We seek to capture the different semantic functions of a given word-sequence by its different contexts, each function giving different probabilities to words in the context. Where T is an occurrence of a word-sequence, let $\boldsymbol{W}$ be the sequence of words around T. Where S is an enumeration of the different ways T contributes to the semantics and Y varies over years, we consider a probability model for $p(Y, S, \boldsymbol{W})$, which by the chain-rule is $p(Y)p(S|Y)p(\boldsymbol{W}|S, Y)$. We assume $\boldsymbol{W}$ is conditionally independent of Y given S, giving $p(Y)p(S|Y)p(\boldsymbol{W}|S)$, and treat $p(\boldsymbol{W}|S)$ as $\prod_i (p(\boldsymbol{W}_i|S)$. Having word probabilities be time-independent given a particular usage firstly reflects the idea that a given concept is accompanied by substantially time-independent vocabulary and secondly drastically reduces the number of parameters that need to be estimated: with 20 time spans and a 2-way usage choice, the word probabilities are conditioned on 2 settings rather than 40.

The parameters are estimated by an Expectation-Maximisation (EM) approach, since the semantic-variant variable, S, is hidden. Space precludes giving all the details, but supposing at each iteration, γ is a table such that for each data point d, and possible S value s, $\gamma[d][s]$ stores $P(S = s|Y = y^d, \boldsymbol{W} = \boldsymbol{w}^d)$, the update formulae work out to be

$$P(S=s|Y=y) = \frac{\sum_d(\text{if } Y^d = y \text{ then } \gamma[d][s] \text{ else } 0)}{\sum_d(\text{if } Y^d = y \text{ then } 1 \text{ else } 0)} \quad P(w|S=s) = \frac{\sum_d(\gamma[d][s] \times freq(w \in \boldsymbol{W}^d))}{\sum_d(\gamma[d][s] \times length(\boldsymbol{W}^d))}$$

Running this EM method on the downloaded data for a particular n-gram infers values for $p(S|Y)$ – inferred usage distributions for each time span. Also approximately 10% was usage-annotated per time-span, giving empirical usage distributions per time span. Figure 1 shows the outcomes.

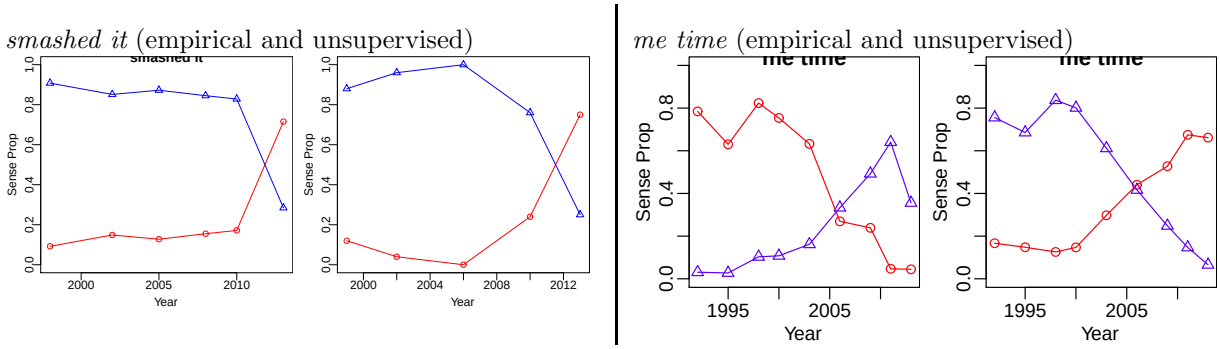

Figure 1: Outcomes on *smashed it* and *me time*, showing empirical sense distributions per time-period in the labelled subset and unsupervised inferred sense distributions per time-period in the entire data set

For *smashed it*, the ∘ line in the empirical plot is for the mwe-usage, and for *me time* it is the △ line, and it has an upward trend. In the unsupervised case, there is an inevitable indeterminacy about which of the semantic option indicators may come to be associated with any given sense. Modulo this the unsupervised and supervised graphs broadly concur. We have looked also at the expressions *biological clock* and *going forward*, finding similar empirical and observed emergence of a recent novel usage.

These are preliminary results. We would like to compare to [Lau et al., 2012] who attempt something comparable for single-word sense induction, contrasting two time-periods, the late 20th century (BNC) and 2007 (ukWac), and we would like to consider other time-stamped corpora [Brants and Franz, 2012, Graff et al., 2007]. A direction for further work is the possible integration of measures of idiomaticity [Biemann and Giesbrecht, 2011].

References

[Biemann and Giesbrecht, 2011] Biemann, C. and Giesbrecht, E., editors (2011). *Proceedings of the Workshop on Distributional Semantics and Compositionality*.

[Brants and Franz, 2012] Brants, T. and Franz, A. (2012). Google books n-grams. ngrams.googlelabs.com.

[Emms, 2013] Emms, M. (2013). Dynamic em in neologism evolution. In *Proceedings of IDEAL 2013*.

[Graff et al., 2007] Graff, D., Kong, J., Chen, K., and Maeda, K. (2007). English gigaword corpus. Linguistic Data Consortium.

[Lau et al., 2012] Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 591–601, Stroudsburg, PA, USA. Association for Computational Linguistics.