Detecting change and emergence for multiword expressions

Martin Emms
Department of Computer Science
Trinity College, Dublin
Ireland
Martin.Emms@tcd.ie

Arun Jayapal
Department of Computer Science
Trinity College, Dublin
Ireland
jayapala@tcd.ie

Abstract

This work looks at a temporal aspect of multiword expressions (MWEs), namely that the behaviour of a given n-gram and its status as a MWE change over time. We propose a model in which context words have particular probabilities given a usage choice for an n-gram, and those usage choices have time dependent probabilities, and we put forward an expectation-maximisation technique for estimating the parameters from data with no annotation of usage choice. For a range of MWE usages of recent coinage, we evaluate whether the technique is able to detect the emerging usage.

1 Introduction

When an n-gram is designated a 'multiword expression', or MWE, its because it possesses properties which are not straightforwardly predictable given the component words of the n-gram – that *red tape* can refer to bureaucratic regulation would be a simple example. A further aspect is that while some *tokens* of the n-gram *type* may be examples of the irregular MWE usage, others may not be – so *red tape* can certainly also be used in a fashion which is transparent relative to its parts. A further aspect is temporal: that tokens of the n-gram can be sought in language samples from different *times*. It seems reasonable to assume that the irregular MWE usage of *red tape* at some time emerged, and was predated by the more transparent usage. This paper concerns the possibility of finding automatic, unsupervised means to detect the emergence of a MWE usage of a given n-gram.

To illustrate further, consider the following examples (these are all taken from the data set on which we worked)

(1)
(a) *the wind lifted his three-car garage and* **smashed it** *to the ground.* (1995)
(a′) *sensational group CEO, totally* **smashed it** *in the BGT (Britain Got Talent)* (2013)
(b) *my schedule gave* **me time** *to get adjusted* (1990)
(b′) *it's important to set time out and enjoy some* **me time** (2013)

(a) and (a′) feature the n-gram *smashed it*. (a) uses the standard destructive sense of *smashed*, and *it* refers to an object undergoing the destructive transformation. In (a′) the n-gram is used differently and is roughly replaceable by 'excelled', a usage not via the standard sense of *smashed*, nor one where *it* refers to any object at all. Where in both (a) and (a′) the n-gram would be regarded as a phrase, (b) and (b′) involving the n-gram *me time* show another possibility. In (b), *me* and *time* are straightforward dependants of *gave*. In (b′), the two words form a noun-phrase, meaning something like 'personal time'. The usage is arguably more acceptable than would be the case with other object pronouns, and if addressed to a particular person, the *me* would refer to the addressee, which is not the usual function of a first-person pronoun.

For *smashed it* and *me time*, the second (primed) example illustrates an irregular usage-variant of the n-gram, whilst the first illustrates a regular usage-variant, and the irregular example is drawn from a later time than the regular usage. Language is a dynamic phenomenon, with the range of ways a given n-gram might contribute subject to change over time, and for these n-grams, it would seem to be the case that the availability of the '*me time*' = '*personal time*' and '*smashed it* = '*excelled*' usage-variants is a relatively recent innovation[1], predated by the regular usage-variants. It seems that in work on multiword ex-

[1]That is to say, recent in British English according to the

pressions, there has been little attention paid to this dynamic aspect, whereby a particular multi-word usage starts to play a role in a language at a particular point in time. Building on earlier work (Emms, 2013), we present some work concerning unsupervised means to detect this. Section 2 describes our data, section 3 our EM-based method and section 4 discusses the results obtained.

2 Data

To investigate such emergence phenomena some kind of time-stamped corpus is required. The approach we took to this was to exploit a search facility that Google has offered for some time – *custom date range* – whereby it is possible to specify a time period for text matching the searched item. To obtain data for a given n-gram, we repeatedly set different year-long time spans and saved the first 100 returned 'hits' as potential examples of the n-gram's use. Each 'hit' has a text snippet and an anchor text for a link to the online source from which the snippet comes. If the text snippet or anchor string contains the n-gram it can furnish an example of its use, and the longer of the two is taken if both feature the n-gram.

A number of n-grams were chosen having the properties that they have an irregular, MWE usage alongside a regular one, with the MWE usage a recent innovation. These were *smashed it*, *me time* (illustrated in (1)) and *going forward*, and *biological clock*, illustrated below.

(c) **Going forward** *from the entrance, you'll come to a large room.* (1995) (2)
(c′) **Going forward** *BJP should engage in people's movements* (2009)
(d) *A* **biological clock** *present in most eukaryotes imposes daily rhythms* (1995)
(d′) *How To Stop Worrying About Your* **Biological Clock** *. . . Pressure to have a baby before 35* (2009)

Alongside the plain movement usage-variant seen in (c), *going forward* has the more opaque usage-variant in which it is roughly replaceable by 'in the future', seen in (c′). Alongside a technical use in biology seen in (d), *biological clock* has come to be used in a wider context to refer to a sense of expiring time within which people may be able to have a child, seen in (d′).

first author's intuitions. It is not easy to find sources to corroborate such intuitions

For each n-gram data was downloaded for successive year-long time-spans from 1990 to 2013, retaining the first 100 hits for each year. For some of the earlier years there are less than 100 hits, but mostly there are more than 100. This gives on the order of 2000 examples for each n-gram, each with a date stamp, but otherwise with no other annotation. See Section 4 for some discussion of this method of obtaining data.

3 Algorithm

For an n-gram with usage variants (as illustrated by (1) and (2)), we take the Bayesian approach that each variant gives different probabilities to the words in its immediate vicinity, as has been done in unsupervised word-sense disambiguation (Manning and Schütze, 2003; de Marneffe and Dupont, 2004). In those approaches, which ignore any temporal dimension, it is also assumed that there are *prior* probabilities on the usage-variants. We bring in language change by having a succession of priors, one for each time period.

To make this more precise, where T is an occurrence of a particular n-gram, with $\boldsymbol{W}$ the sequence of words around T, let Y represent its time-stamp. If we suppose there are k different usage-variants of the n-gram, we simply model this with a discrete variable S which can take on k values. So S can be thought of as ranging over positions in an enumeration of the different ways that the n-gram can contribute to the semantics. With these variables we can say that we are considering a probability model for $p(Y, S, \boldsymbol{W})$. Applying the chain-rule this may be re-expresssed without loss of generality as $p(Y)p(S|Y)p(\boldsymbol{W}|S, Y)$. We then make some assumptions: (i) that $\boldsymbol{W}$ is conditionally independent of Y given S, so $p(\boldsymbol{W}|S, Y) = p(\boldsymbol{W}|S)$, (ii) that $p(\boldsymbol{W}|S)$ may be treated as $\prod_i (p(\boldsymbol{W}_i|S)$, and (iii) that $p(Y)$ is uniform. This then gives

$$p(Y, S, \boldsymbol{W}) = p(Y)p(S|Y) \prod_i (p(\boldsymbol{W}_i|S) \quad (3)$$

The term $p(S|Y)$ directly models the fact that a usage variant can vary its likelihood over time, possibly having zero probability on some early range of times. While (i) make context words and times indepedent *given* a usage variant, context words are still time-dependent: the sum $\sum_S [p(S|Y)p(\boldsymbol{W}|S)]$ varies with time Y due to

$p(S|Y)$. Assumption (i) reflects a plausibile idea that given a concept being conveyed, the expected accompanying vocabulary is substantially time-independent. Moreover (i) drastically reduces the number of parameters 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 of the model in (3) have to be estimated from data which is labelled only for time – the usage-variant variable is a *hidden* variable – and we tackle this with an EM procedure (Dempster et al., 1977). Space precludes giving the derivations of the update formulae but in outline there is an iteration of an E and an M step, as follows:

(**E** step) *based on current parameters, a table,* γ*, is populated, such that for each data point* d*, and possible* S *value* s*,* $\gamma[d][s]$ *stores* $P(S = s|Y = y^d, \boldsymbol{W} = \boldsymbol{w}^d)$.

(**M** step) *based on* γ*, fresh parameter values are re-estimated according to:*

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

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

These updates can be shown to increase the data probability, where the usage variable S is summed-out.

4 Results and Discussion

Running the above-outlined EM procedure on the downloaded data for a particular n-gram generates unsupervised estimates for $p(S|Y)$ – inferred usage distributions for each time span. To obtain a reference with which to compare these inferred distributions, approximately 10% of the data per time-span was manually annotated and used to give simple relative-frequency estimates of $p(S|Y)$ – which we will call empirical estimates. Although the data was downloaded for year-long time spans, it was decided to group the data into successive spans of 3 year duration. This was to make the empirical $p(S|Y)$ less brittle as they are otherwise based on too small a quantity of data.

Figure 1 shows the outcomes, as usage-variant probabilities in a succession of time spans, both the empirical estimates obtained on a subset, and the unsupervised estimates obtained on all the data. The EM method can seek any number of usage variants, and the results show the case where 2 variants were sought. Where the manually annotated subset used more variants these were grouped to facilitate a comparison.

For *smashed it*, *biological clock* and *going forward*, the $\circ$ line in the empirical plot is for the MWE usage, and for *me time* it is the $\triangle$ line, and it has an upward trend. In the unsupervised case, there is inevitable indeterminacy about which S values may come to be associated with any objectively real usage. Modulo this the unsupervised and supervised graphs broadly concur.

One can also inspect the context-words which come to have high probability in one semantic variant relative to their probability in another. For example, for *smashed it*, for the semantic usage which is inferred to have an increasing probability in recent years, a selection from the most favoured tokens includes *!!, guys, really, completely, They, !*, whilst for the other usage they include *smithereens, bits, bottle, onto, phone*. For *biological clock*, a similar exercise gives for the apparently increasing usage, tokens such as *Ticks, Ticking?, Health, Fertility* and for the other usage *running, 24-hour, controlled, mammalian, mechanisms*. These associations would seem to be consistent with the inferred semantic-usages being in broad correspondence with the annotated usages.

As noted in section 2, as a means to obtain data on relatively recent n-gram usages, we used the *custom date range* search facility of Google. One of the issues with such data is the potential for the time-stamping (inferred by Google) to be innaccurate. Though its not possible to exhaustively verify the time-stamping, some inspection was done, which revealed that although there are some cases of documents which were incorrectly stamped, this was tolerably infrequent. Then there is the question of the representativeness of the sample obtained. The mechanism we used gives the first 100 from the at most 1000 'hits' which Google will return from amongst all index documents which match the n-gram and the date range, so an uncontrollable factor is the ranking mechanism according to which these hits are selected and ordered. The fact that the empirical usage distributions accord reasonably well with prior intuition is a modest indicator that the data is not unusably unrepresentative. One could also argue that for an initial test of the algorithms it suffices for the methods to recover an apparent trend

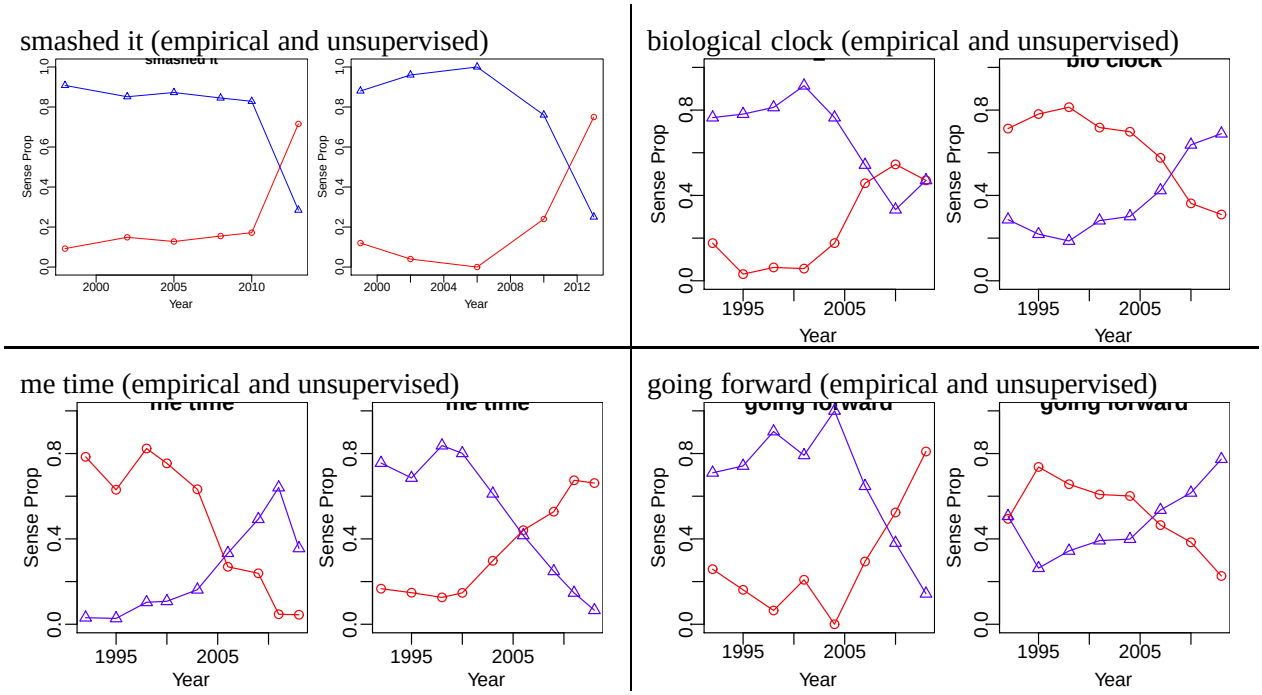

Figure 1: For each n-gram the plots show the empirical usage-variant distributions per time-period in the labelled subset and unsupervised usage-varaint distributions per time-period in the entire data set

in the downloaded data, even if the data is unrepresentative. This being said, one direction for further work will be to consider other sources of time-stamped language use, such as the Google n-grams corpus (Brants and Franz, 2012), or various newswire corpora (Graff et al., 2007).

There does not seem to have been that much work on unsupervised means to identify emergence of new usage of a given expression – there is more work which groups all tokens of a type together and uses change of context words to indicate an evolving single meaning (Sagi et al., 2008; Gulordava and Baroni, 2011). Lau et al. (2012) though they do not address MWEs do look at the emergence of new word senses, applying a word-sense induction technique. Their testing was between two corpora taken to represent two different time periods, the BNC and ukWac corpus, taken to represent the late 20th century and 2007, respectively, and they reported promising results on 5 words. The unsupervised method they used is based on a Hierarchical Dirichlet Process model (Yao and Van Durme, 2011), and a direction for future work will be a closer comparison of the algorithm presented here to that algorithm and other related LDA-based methods in word sense induction (Brody and Lapata, 2009). Also the bag-of-tokens model of the context words which we adopted is a very simple one, and we wish to consider more sophisticated models involving for example part-of-tagging or syntactic structures.

The results are indicative at least that MWE usage of an n-gram can be detected by unsupervised means to be preceded by the other usages of the n-gram. There has been some work on algorithms which seek to quantify the degree of compositionality of particular n-grams (Maldonado-Guerra and Emms, 2011; Biemann and Giesbrecht, 2011) and it is hoped in future work to consider the possible integration of some of these techniques with those reported here. For a given n-gram, it would be interesting to know if the collection of its occurrences which the techniques of the current paper suggest to belong to a more recently emerging usage, are also a corpus of occurrences relative to which a compositionality measure would report the n-gram as being of low compositionality, and conversely for the apparently less recent usage.

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.

References

Chris Biemann and Eugenie Giesbrecht, editors. 2011. *Proceedings of the Workshop on Distributional Semantics and Compositionality*.

Thorsten Brants and Alex Franz. 2012. Google books n-grams. ngrams.googlelabs.com.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *EACL 09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catharine de Marneffe and Pierre Dupont. 2004. Comparative study of statistical word sense discrimination. In Gérald Purnelle, Cédric Fairon, and Anne Dister, editors, *Proceedings of JADT 2004 7th International Conference on the Statistical Analysis of Textual Data*, pages 270–281. UCL Presses Universitaire de Louvain.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statistical Society*, B 39:1–38.

Martin Emms. 2013. Dynamic EM in neologism evolution. In Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li, and Xin Yao, editors, *Proceedings of IDEAL 2013*, volume 8206 of *Lecture Notes in Computer Science*, pages 286–293. Springer.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English gigaword corpus. Linguistic Data Consortium.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 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.

Alfredo Maldonado-Guerra and Martin Emms. 2011. Measuring the compositionality of collocations via word co-occurrence vectors: Shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 48–53, Portland, Oregon, USA, June. Association for Computational Linguistics.

Christopher Manning and Hinrich Schütze, 2003. *Foundations of Statistical Language Processing*, chapter Word Sense Disambiguation, pages 229–264. MIT Press, 6 edition.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2008. Tracing semantic change with latent semantic analysis. In *Proceedings of ICEHL 2008*.

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14. Association for Computational Linguistics.