Dynamic EM in Neologism Evolution

Martin Emms

Dept. of Computer Science, Trinity College, Dublin, Ireland
www.scss.tcd.ie/Martin.Emms

Abstract. Research on unsupervised word sense discrimination typically ignores a notable dynamic aspect, whereby the prevalence of a word sense varies over time, to the point that a given word (such as 'tweet') can acquire a new usage alongside a pre-existing one (such as 'a Twitter post' alongside 'a bird noise'). This work applies unsupervised methods to text collections within which such *neologisms* can reasonably be expected to occur. We propose a probabilistic model which conditions words on senses, and senses on times and an EM method to learn the parameters of the model using data from which sense labels have been deleted. This is contrasted with a static model with no time dependency. We show qualitatively that the learned and the observed time-dependent sense distributions resemble each other closely, and quantitatively that the learned dynamic model achieves a higher tagging accuracy (82.4%) than the learned static model does (76.1%).

Keywords: neologism sense EM

1 Introduction

Language is subject to continuous change, one aspect of which is lexical: an existing word can acquire a new usage. Arguably content that has been relatively recently authored for access via the internet is relatively rich in such *neologisms.*

For example, a post to *Twitter* is known as a *tweet.* This is a new usage, pre-dated by the use of *tweet* to refer to a particular kind of bird noise. Related to this *Twitter*-related noun usage of *tweet* there is also *tweeted* and *tweeter.* Before the era of Twitter, the pre-eminent usage of *tweeter* was to designate a particular piece of audio hardware.

The n-gram *totes* has acquired a usage where it substitutes for *totally,* (as in *I was totes steamed*). This is pre-dated by several other usages, one where it is the plural of *tote*, itself ambiguous between designating a certain kind of bag or designating a man who takes bets at a race course.

The n-gram *bricked* has acquired a usage meaning roughly to render a piece of equipment, often a phone, entirely unresponsive. This is pre-dated by a usage referring to a construction process involving bricks.

Whilst there are many interesting questions as to the mechanisms at work in creating these neologisms, this article is concerned with the simple fact that such new usages add to the ambiguity-resolution problem that is a feature of many NLP tasks, such as information retrieval or machine translation.

To illustrate the fact that some of these neologisms are not well handled, we sought their translation into German via Google Translate[1]:

English	German (via Google Translate)
he is a regular tweeter	*er ist ein regelmaessiger Hochtoener*
he has bricked my phone	*er hat mein Handy zugemauert*

The chosen translation of *tweeter*, namely *Hochtoener*, would have been appropriate to its audio hardware sense. It is not appropriate to its Twitter-related sense. Similarly the chosen translation of *bricked*, namely *zugemauert*, would have been appropriate to a construction setting, but is not appropriate to the context of rendering a phone inert.[2]

The terms *unsupervised sense disambiguation* or *word sense discrimination* are often used to describe a process which takes as input a corpus of examples of a word's use – unlabelled with any sense information – and returns a partitioning of the examples into groups, so that all members of a group exhibit a particular sense or usage [9, 7, 5, 2]. Often in such systems, each cell of the partition is associated with parameters, and via these for any given example of the word's use, a number can be derived which can be interpreted as strength of membership in that cell. For example, in Schütze's work [9], each cell of the partitition is associated with a centroid vector, and distance from this indicates strength of membership. The outputs of such word sense discrimination systems can have practical value. For example, for a probabilistically trained MT system, one might be able to then recognise that in a sentence to be translated, a given word belongs only weakly to any of the cells of the partition for that word in the training corpus for the system and so is likely to be poorly translated [11]. There is also work reporting improved performance in information retrieval [10].

The work to be presented below is in the area of word sense discrimination, but its innovative aspect is that we seek to address the above-noted fact that language is a moving target. One can expect that the prevalence of use of a particular sense varies with time. For the kind of neologisms mentioned above – new usages of an existing word – it should be the case that prior to a particular point in time, usage with the new sense is non-existent. For example, Twitter was founded in 2006, so that the Twitter-related sense of *tweet* should be absent from text authored prior, or substantially prior, to that date.

Thus our corpus of word occurrences will span a stretch of time. If there is change in the prevalence of use of a particular sense over that time period, possibly even the emergence of a particular sense at a particular time, the question we wish to address is whether that can be detected by an unsupervised method, and whether one particular model is better than another to accomplish this.

We will propose a probabilistic model of word use which conditions words on senses, and senses on times, and we will propose a *dynamic EM* method for

[1] Executed May 2013.

[2] Some internet search and consultation with German-speaking colleagues suggest that *gebrickt* would be the German equivalent.

estimating the parameters of the model from data in which times and words are visible, but senses hidden. This will be contrasted with a static variant with senses unconditioned on times. We will give qualitative evidence that the method is able to infer changes in the prevalence of a sense over time, and quantitative evidence that the dynamic variant out performs the static variant, achieving an accuracy of 82.4% vs 76.1%, when the partitions are mapped to manual assigned senses.

2 Dynamic and Static Models

The following snippet of text, dating from 2001, is an example of the construction-related sense of *bricked*:

> *. . . In 1611 she was* **bricked** *into one of the rooms . . .*

and the next snippet, dating from 2011, illustrates the 'render inert' sense:

> *I've tried to flash a custom ROM and now I think I've* **bricked** *my phone*

Where T is an occurrence of a target ambigous word, let $\boldsymbol{W}$ be a sequence whose first l elements are the l words to the left of T and whose last r elements are the r words to the right of T. Assume the document in which T features was authored in a particular year, represented by variable Y. Let S range over available senses for target ambiguous term. We assume that the occurrence T uses a particular sense. If we suppose there are three available senses, with $S = 1$ for the construction-related sense and $S = 2$ for the phone-related sesnse, then choosing $l = r = 5$, the two examples above might be represented thus:[3]

$$Y = 2001,\ S = 1,\ \boldsymbol{W} = \langle L, In, 1611, she, was, into, one, of, the, rooms\rangle$$
$$Y = 2011,\ S = 2,\ \boldsymbol{W} = \langle and, now, I, think, I've, my, phone, R, R, R\rangle$$

We wish to consider possible probability models for combinations of the variables Y, S and $\boldsymbol{W}$. Without loss of generality, using the chain rule, we have

$$p(Y, S, \boldsymbol{W}) = p(Y) \times p(S|Y) \times p(\boldsymbol{W}|S, Y)$$

Considering the last term in the above, we will assume that $\boldsymbol{W}$ is conditionally independent of Y given S, that is, $p(\boldsymbol{W}|S, Y) = p(\boldsymbol{W}|S)$, so that

$$p(Y, S, \boldsymbol{W}) = p(Y) \times p(S|Y) \times p(\boldsymbol{W}|S) \tag{1}$$

Whilst this is a simplifying assumption, there is some plausibility to the idea that given the concept that is being conveyed, there is some substantially time-independent aspect of the expected accompanying vocabulary. The assumption drastically reduces the number of parameters that need to be estimated: with

[3] The context is padded with L and R if its not wide enough.

a 10-year time span and a 3-way sense ambiguity, the word probabilities are conditioned on 3 settings rather than 30. We further treat $p(\boldsymbol{W}|S)$ as if $\boldsymbol{W}$ represents $|\boldsymbol{W}|$ trials of a single multinomial variable, whose values are the different possible words, hence $p(\boldsymbol{W}|S) = \prod_i (p(\boldsymbol{W}_i|S)$ (essentially a so-called unigram language model [4]).

The second term in (1) directly expresses the idea that the prevalence of a given sense can vary with the year. For example, on a-priori grounds, we can be sure that the Twitter-relatd sense of *tweet* has zero probability for $Y = 1972$.

Whilst (1) gives a *dynamic* model, the further simplifying assumption that the sense is *independent* of the time, $p(S|Y) = p(S)$, gives the *static* model (2)

$$p(Y, S, \boldsymbol{W}) = p(Y) \times p(S) \times p(\boldsymbol{W}|S) \tag{2}$$

If all mentions of the time Y are removed from the data and (2) then the static model is essentially the probabilistic model considered in several works ([3, 5]).

If the probabilities mentioned on the righthand sides of the equations (1) and (2) are known then for an instance with given values for the Y and $\boldsymbol{W}$, the instance can be classified (resp. assigned a probability) by choosing the value in $val(S)$ that maximises (resp. taking the sum over all values in $val(S)$ of) the joint probability $p(Y = y, S, \boldsymbol{W} = \boldsymbol{w})$.[4]

In our training data, the sense variable S will be *hidden*, and to estimate the parameters of the proposed model from this data we propose to use an Expectation-Maximisation (EM) approach to obtain values for the parameters – for what is essentially the *static* model of (2), this is discussed in [3, 5]. The development below is based on the idea of repeatedly calculating expected completions of incomplete data, and then deriving new parameters from these by maximum likelihood estimation. [6] is a good exposition of this perspective on EM.

Let $\mathcal{T} = \ldots (Y^d, \boldsymbol{W}^d) \ldots$, be a corpus of instances: note the sense S is not present. Then in outline the training algorithm consists in iterations of the following pair of steps, mapping the estimates $\theta_n(p(Y), p(S|Y), p(\boldsymbol{W}|S))$ to new estimates $\theta_{n+1}(p(Y), p(S|Y), p(\boldsymbol{W}|S))$.

(Exp) *generate a virtual corpus of disambiguated instances by treating each training instance* $(Y^d, \boldsymbol{W}^d)$ *as standing for* all *possible completions with a sense,* $(Y^d, S, \boldsymbol{W}^d)$*, weighting* each *by its conditional probability* $P(S|Y^d, \boldsymbol{W}^d)$*, under current probalities* $\theta_n(p(Y), p(S|Y), p(\boldsymbol{W}|S))$

(Max) *apply maximum likelihood estimation to the virtual corpus to derive new estimates* $\theta_{n+1}(p(Y), p(S|Y), p(\boldsymbol{W}|S))$*.*

Concretely, where D is the size of the training data, let γ be a table of size $D \times |val(S)|$ to store for each data point d the conditional probabilities $P(S = s|Y = y^d, \boldsymbol{W} = \boldsymbol{w}^d)$ for each $s \in val(S)$. For each d, $\gamma[d][s]$ is calculated according to

[4] For any variable X with discrete values, let $val(X)$ be its set of possible values.

$$\gamma[d][s] = \frac{P(Y = y^d, S = s, \boldsymbol{W} = \boldsymbol{w}^d)}{\sum_{s' \in val(S)} P(Y = y^d, S = s', \boldsymbol{W} = \boldsymbol{w}^d)}$$

Supposing γ has been filled by a traversal of the training data $\{..., (Y^d, \boldsymbol{W}^d), ...\}$. For the dynamic model of (1) the re-estimation formulae for the (Max) step are as follows

For each $y \in val(Y)$, and $s \in val(S)$

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

For each $s \in val(S)$

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

For the static model of (2), the re-estimation of $P(S)$ will be via

$$P(S = s) = \frac{\sum_d (\gamma[d][s])}{D}$$

3 Experiments

As mentioned earlier *bricked* seems to have acquired a sense relating to rendering a technical device (often a phone) inert, in addition to others senses, principally a building-related sense. To test the approach for *bricked*, a corpus of time-stamped occurrences was required. This was obtained using a facility that Google has offered for some time in which it is possible to specify a time period for searched documents. Eleven year-long search periods were used, (1/1/2001-31/12/2002) . . . (1/1/2011-31/12/2012).

For each of these time periods, the search item was set to *bricked* and the first 20 hits retained. These were then used to give time specific occurrences of the search item, and a window of 5 words to left and right was used to define $\boldsymbol{W}$, giving 220 occurrences, 20 for each year. For later evaluation purposes these were manually inspected and labelled with one of three sense tags: *WALLS*, for the construction-related sense, *PHONE* for the render-inert sense, and *OTHER* of any other sense. Two examples of this data were given at the beginning of Sect. 2.

Based on the manually assigned sense labels, Fig. 1 shows for each year the relative proportions of occurences for the WALLS and PHONE senses of *bricked*. It is clear from this plot that, at least for the sample corpus obtained, there is a substantial shift in the proportion of occurrences accounted for by *WALLS* and *PHONE* senses. In 2001, *WALLS* predominates, and *PHONE* is negligible, whilst by 2011 the situation is reversed, and *PHONE* predominates and *WALLS*

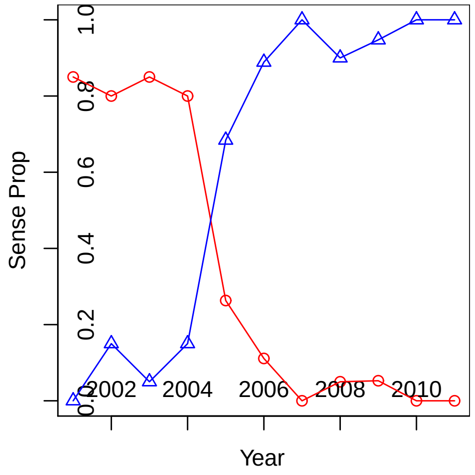

Fig. 1. Observed per-year proportions of two senses of *bricked* from 2001 to 2011 (*WALLS* shown as circles, *PHONE* shown as triangles)

is negligible, with the cross-over happening somewhere between the beginning of 2004 and the end of 2005. Combining all years together, the relative proportions are *WALLS* (0.357) and *PHONE* (0.6)

We ran the EM training algorithm on this data set, assuming the dynamic model of (1). Recall that in this training the sense values are *hidden*. When the sense variable is summed-out, the EM training increased the likelihood of the data, as it theoretically should. Of more interest though are the conditional sense-given-year probabilities after the training process has converged and these are shown in Fig. 2. Qualitatively one can see that the obtained values resemble quite closely the empirical sense distributions which were shown in Fig. 1.

Recall that although the unsupervised EM training operated on an unlabelled version of the data, there are manually assigned labels available for the data. If one defines a mapping, σ, from the sense identifiers of the model, $\{1, 2, 3\}$, to the sense identifiers in the data, $\{WALLS, PHONE, OTHER\}$, one can evaluate the accuracy of the labelling produced by the model at the end of the unsupervised EM training, modulo the mapping σ. As there are many possible mappings, following the methodology of [7], we report the highest accuracy under all mappings. The accuracy obtained in this way is 82.4%. When instead the *static* model is used from (2), the trained model's accuracy is somewhat lower: 76.1%.

4 Discussion and Conclusions

We have given some preliminary evidence concerning the possibility to use an EM method to infer senses on data where the sense distributions are intuitively

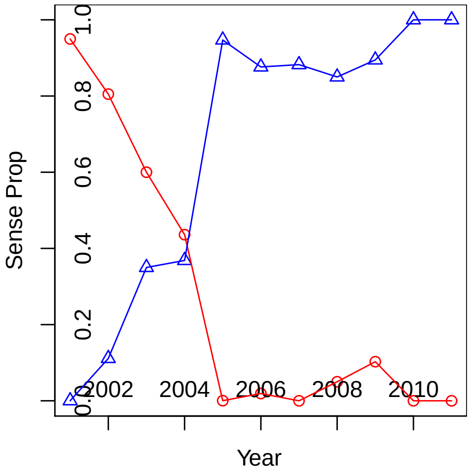

Fig. 2. Values of $P(S|Y)$ across the years after EM training using the dynamic model.

time dependent. A model incorporating this time dependence was shown to outperform a static model. As far as we are aware, this has not been done before.

Work on *Dynamic Topic Models* [1] takes a somewhat analogous approach, with *topic* (such as 'Atomic Physics' or 'Neuroscience') as a hidden, time-dependent variable, and probabilities on a *document*'s words which are topic and time dependent. Besides being applied to document topics rather than word senses, the method used – variational approximation – differs substantially to that used here. It remains for future work to make a more detailed comparison. A preliminary observation would be that the additional *time*-dependency of word probabilities for topics seems more natural than it is for word-senses.

Although the current work is concerned with an interaction between senses and times, it still treats senses as somehow *eternal*: abstractly for a given span of time and a given word, there is a fixed set of available senses and what varies is the probability for a sense over time. [8] reports work in historical linguistics on tracing semantic change which uses the context vectors of [9] to analyse changes in word senses themselves. For example, in comparison to hundreds of years ago, *dog* (resp. *deer*) has come to have a much broader (resp. narrower) sense, and they claim to be able to map this to changes in the geometry of the set of vectors for a word over time. It remains for future work to consider more closely the relation between the work of [8] and the problem addressed here.

This work was a preliminary investigation. One issue needing attention is the use of Google's time-specific search facility. For the lexical item looked at, the time-stamping seemed seldom glaringly inaccurate, but for others that were considered, it seemed so. Also, although we made the simplifying assumption that the vocabulary probabilities were independent of the time, given the sense, it is easy to imagine situations where this is not going to be true. One direction for future work would be to investigate outcomes in which this simplifying assumption is not made, but the increased sophistication of such a model certainly suggests that it would demand a far larger corpus.

Besides attempting to use the inferred model to assign sense-group labels to words, one could instead use the approach to attempt to simply recognise that over a particular span of time there appears to be a newly emerging sense, or neologism.

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

1. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Cohen, W., Moore, A. (eds.) ICML 06: Proceedings of the 23rd International Conference on Machine Learning. pp. 113–120. ACM Press, New York, NY, USA (2006)
2. Maldonado-Guerra, A., Emms, M.: First-order and second-order context representations: geometrical considerations and performance in word-sense disambiguation and discrimination. In: Dister, A., Dominique Longrée, G.P. (eds.) Proceeding of JADT 11th International Conference on the Statistical Analysis of Textual Data. pp. 676–686. LASLA (2012)
3. Manning, C., Schütze, H.: Foundations of Statistical Language Processing, chap. Word Sense Disambiguation, pp. 229–264. MIT Press, 6 edn. (2003)
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, chap. Language models for information retrieval. Cambridge University Press (2009)
5. de Marneffe, M.C., Dupont, P.: Comparative study of statistical word sense discrimination. In: Purnelle, G., Fairon, C., Dister, A. (eds.) Proceedings of JADT 2004 7th International Conference on the Statistical Analysis of Textual Data. pp. 270–281. UCL Presses Universitaire de Louvain (2004)
6. Prescher, D.: A tutorial on the expectation-maximization algorithm including maximum-likelihood estimation and em training of probabilistic context-free grammars. Computing Research Repository (2004)
7. Purandare, A., Pedersen, T.: Word sense discrimination by clustering contexts in vector and similarity spaces. In: Ng, H.T., Riloff, E. (eds.) Proceedings of CoNLL-2004. pp. 41–48. Boston, MA, USA (2004)
8. Sagi, E., Kaufmann, S., Clark, B.: Tracing semantic change with latent semantic analysis. In: Allan, K., Robinson, J.A. (eds.) Current Methods in Historical Semantics, pp. 161–183. Mouton de Gruyter, Berlin (2012)
9. Schütze, H.: Automatic word sense discrimination. Computational Linguistics 24(1), 97–123 (1998)
10. Véronis, J.: Hyperlex: lexical cartography for information retrieval. Computer Speech and Language 18(3), 223–252 (2004)
11. Vickrey, D., Biewald, L., Teyssier, M., Koller, D.: Word-sense disambiguation for machine translation. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 771–778. HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005)