An unsupervised EM method to infer time variation in sense probabilities

Martin Emms
Dept of Computer Science
Trinity College, Dublin
Ireland
martin.emms@cs.tcd.ie

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

Abstract

The inventory of senses for a given word changes over time – *tweet* has gained the ‘Twitter post’ sense only relatively recently and this paper addresses the problem of the computational detection of such change. We propose a generative model which conditions context words on a target expression’s sense and conditions the sense choice on the *time* of writing. We develop an EM algorithm to estimate the parameters from raw time-stamped n-gram data with no sense annotation. We are able to demonstrate the inference of parameters which plausibly reflect the objective dynamics of sense use frequencies, in particular the emergence of a new sense.

1 Introduction

Many words are ambiguous so that a simple count of a word frequency is really a sum of several word-*sense* frequencies. It is also the case that over time the inventory of senses possessed by a word *changes* and there must be a point in time where a given sense first came into use: the ‘Twitter post’ sense of *tweet* is an example of a relatively recent new arrival. Fig 1 summarises the situation abstractly, with the solid black-line a plot of simple word frequency over time (as might easily be produced the Google n-gram viewer), and the dashed lines a hypothetical decomposition into two senses, with one of the senses first emerging around 1970. Essentially the aim of the paper is to propose a method which can carry out this kind of decomposition and then use it to detect the time of emergence of a novel sense: we have marked this point C_0 in the sense plot.

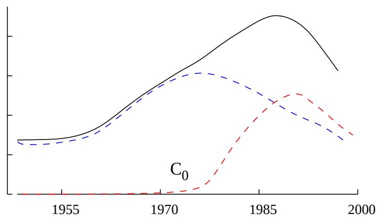


Figure 1: Word frequency (solid line) and sense frequencies (dashed lines).

As lexical information is central to so many NLP tasks, means to automatically identify *changes* to the required information could be useful. For example, if an SMT system trained from aligned corpora from particular times is to be applied to text from different times it could be of use to know whether there have been sense changes, perhaps identifying which occurrences can be anticipated to be poorly translated. Overlooking this is perhaps what underlies the following case of poor translation fron English to Tamil via Google translate:

S1. With Clara, however, his brow cleared, and he was gay again (from ‘sons and lovers’ by D.H. Lawrence 1931:)

T1. கிளாரா ஆயினும் அவரது புருவம் அகற்றப்படும் மற்றும் அவர் மீண்டும் ஓரினச்சேர்க்கையாளர்

L1. Kiḷārā, āyiṉum, avaratu puruvam akaṟṟappaṭum, maṟṟum avar mīṇṭum ōriṉaccērkkaiyāḷar

The English original, S1, comes from 1931, at which time *gay* simply meant ‘happy’. In recent decades it has overwhelmingly gained the sense ‘homosexual’. T1 and L1 are the Tamil translation and its transliteration where the word is translated as *ōriṉaccērkkaiyāḷar*, which has the ‘homosexual’ sense.

We will propose a relatively simple probabilistic model which conditions words in a target’s context on that target’s sense and conditions senses on times, which incorporates an independence assumption that the context words are independent of time given the target’s sense. We use the Google n-gram data set (Michel et al., 2011) which provides time-stamped data but *no* sense information and develop an EM algorithm to estimate

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 89–94,
Trivandrum, India. December 2015. ©2015 NLP Association of India (NLPAI)

the model's parameters in an unsupervised fashion from this data. We will show that the algorithm is able to provide an accurate date of sense emergence (true positives), and also to detect the absence of sense emergence when appropriate (true negatives). We make some points also concerning the difficulties in the evaluation of such a sense-emergence system.

2 A model with time-dependent senses

It is intuitive that the distribution of words in the context of given target word is not independent of the sense of that target and this leads to approaches to unsupervised word-sense disambiguation which seek to model that dependency (Manning and Schütze, 2003). These approaches also include a *prior* on the different senses. We essentially generalise this by having a *succession of priors*, one for each time period.

To make this more precise, assume a data item d represents a particular occurrence of a target expression T, and let $\vec{W}$ be the sequence of words in a window around T and let Y be its time-stamp. Suppose there are k different senses of T, modeled with a discrete variable S. A *complete* data item is to be thought of as providing values for Y, S and $\vec{W}$; raw data will be *incomplete* concerning S, the sense. $p(Y, S, \vec{W})$ may factorised as $p(Y)p(S|Y)p(\vec{W}|S, Y)$ without loss of generality. We then make some independence assumptions: (i) that conditioned on S the words $\vec{W}$ are independent of Y, so $p(\vec{W}|S, Y) = p(\vec{W}|S)$ and (ii) that conditioned on S the words are independent of each other, so $p(\vec{W}|S) = \prod_i (p(\vec{W}_i|S)$. This gives equation (1):

$$p(Y, S, \vec{W}) = p(Y) \times p(S|Y) \times \prod_i p(W_i|S) \quad (1)$$

The term $p(S|Y)$ directly models the fact that a sense's likelihood can vary over time, possibly having zero probability on some early range of times and thus representing the non-availability of that sense of target T at those times. Assumption (i) reflects a plausible idea that given a concept being conveyed, the expected accompanying vocabulary is at least substantially time-independent. It also drastically reduces the number of parameters to be estimated. The parameters of the model in (1) have to be estimated from data in which the values of the sense variable S are not given. We tackle this unsupervised estimation problem via an EM procedure (Dempster et al., 1977), though Gibbs sampling could be an alternative. We iterate between an E and an M step, as follows[1]:

(E) for each data item d and each possible value s of S, determine the conditional probability of $S = s$ given Y^d and $\vec{W}^d$, under current parameters – call this $\gamma^d(s)$.

(M) use the $\gamma^d(s)$ valus to derive new estimates of parameters via the update formulae

$$\begin{aligned} P(S = s|Y = y) &= \frac{\sum_{d:Y^d=y}[\gamma^d(s)]}{\sum_{d:Y^d=y}[1]} \\ P(w|S = s) &= \frac{\sum_d(\gamma^d(s) \times \#(w, \vec{W}^d))}{\sum_d(\gamma^d(s) \times length(\vec{W}^d))} \end{aligned}$$

2.1 Evaluation possibilities

As others have noted, the evaluation of the inferred sense dynamics produced by such a system is a challenge: large-scale, sense-labelled, diachronic corpora to serve as a gold-standard do not exist (Cook et al., 2014). So without doing large scale manual annotation, the full detail of the inferred sense dynamics cannot be straightforwardly evaluated.

However, concerning just the times of sense *emergence* (the point C_0 in Fig.1) the prospects are somewhat better. First of all, for *recent* lexical innovations, native speakers are often confident that they can judge that it is indeed recent (see section 4). A speaker's intuition that a sense emerged recently could serve as prima facie evidence against a system verdict outcome which finds no such recent emergence.

In pursuit of more objectivity and to consider innovations that are less recent, it is natural to consider dictionaries. For a given form/meaning pairing, an historically oriented dictionary (eg. the Oxford English Dictionary (OED)) will strive to include the very earliest citation that has been discovered – call this D_0^c. We will use D_0^c as a *lower* bound on the true emergence date C_0, which seems reasonable. The results section 3 will show a couple of examples wherre D_0^c is considerably earlier than C_0, the time at which which the use steadily starts to grow in use. Form/meaning pairings also make it into dictionaries at some particular time, so close inspection of a succession of dictionaries, though labour intensive, can give a

[1] $P(Y)$ is not estimated can be easily shown not to be needed for the other estimates.

date of its first inclusion – call this D^i_0. Some researchers have attempted this (see section 4), though in what follows we will not.

There is a further option to establish an emergence date for a novel sense of a target T. If there are words which it is intuitive to expect in the vicinity of T in the novel sense, and not in other senses, then one would expect the probability of seeing these words in T's context to start to climb at a particular time. For example, *mouse* has come to have a sense referring to a computer pointing device, in which usage it is intuitive to expect words like *click*, *button*, *pointer* and *drag* in it's context. For any word w and target T let us define $track_T(w)$ to be the sequence of its per-year probabilities of occurrence in the context of T. If when $track_{mouse}(w)$ is plotted for the above words, they all show a sharp increase at the *same* time point, this is good evidence that this is the emergence time of the novel sense. In the results section 3 we will use such 'tracks'-based dates as a target against which to compare any apparent inflection point of an inferred $P(S|Y)$ parameter – the righthand plot in Figure 2(a) is an example.

3 Experiments

The experiments reported here use the Google 5-gram dataset (Michel et al., 2011). This is a data set released by Google giving per-year counts of 5-grams in their digitized books holdings. From the entire 5-gram data-set it is possible to pull for a given target word T, a corpus of time-stamped 5-grams (with counts) containing T. As a diachronic corpus this data set has size and time range advantages. For example, prior work (Emms, 2013; Emms and Jayapal, 2014) used text snippets from search results obtained using Google's date specific search facility, giving a time range of 1990-2013 and about *100* examples per year. The 5-gram data reaches far further back with a far larger amount of data per year (see Table 1). For example for the target *mouse* there is on average over *15000* 5-gram entries featuring it per year. Its largest seeming disadvantage is that each 5-gram for a target gives a context of no more than 4 words.

Each line of the Google 5-gram data gives a year-specific *count* for the particular 5-gram. The EM presentation in section 2 assumed data items represented *single* target occurrences. In using the 5-gram data we effectively treat each 5-gram data entry as representing n separate tokens of T contributing to no other 5-gram counts: the data set makes it impossible to know to what extent any original token of T has contributed to several different 5-gram counts. This involves changing the M step to

$$P(S = s|Y = y) = \frac{\sum_{d:Y^d=y}[\gamma^d(s)\times n^d]}{\sum_{d:Y^d=y}[n^d]}$$

$$P(w|S = s) = \frac{\sum_d(\gamma^d(s)\times n^d\times\#(w,\vec{W}^d))}{\sum_d(\gamma^d(s)\times n^d\times length(\vec{W}^d))}$$

Concerning initialisation, for an experiment on a target T having a corpus of occurrences $corp$, we initialise $P(w|S)$ to $(1-\lambda)P_{corp} + \lambda P_{ran}$, where P_{corp} are the word probabilities in $corp$, P_{ran} is a random word distribution and λ is a mixing proportion, here set to 10^{-5}. Also initially the per-year sense distributions $P(S|Y)$ values are set to the same as each other. These start values thus are very far from representing the senses as being drastically different to each other or having any time variation at all.

Two sets of targets where chosen, a first set {*strike*, *mouse*, *boot*, *compile*, *surf* } of words known to exhibit sense emergence and a second set { *ostensible*, *present* } of words where there is not an expectation of an emergent sense. The two sets together give both positive and negative tests for the algorithm. Table 1 lists these words and for the first set gives an indication of the emergent senses. The next two columns give two kinds of reference dating information for the emergence of these senses. The 'OED-first' column gives the first citation date according to the OED, a lower bound for any inferred emergence date. The 'tracks date' gives an emergence date that is apparent from the 'tracks' plots for words that are intuitively associated with the emergent sense (see section 2.1). The final 'EM date' column is the emergence date inferred when the EM algorithm was run.

Figure 2(a-d) depicts various aspects of the outcomes on some of the test items. The leftmost plot in (a-c) shows the EM-inferred $P(S|Y)$ parameters for different targets T, and the rightmost plot shows some $tracks_T(w)$ plots (see section 2.1), for some words thought especially associated with the novel sensse – these track the probabilities of these words in the n-grams for the target and are the basis for the 'tracks date' column in Table 1.

mouse Figure 2(a): the algorithm was run for 3 sense variants on data from 1950 to 2010, and in the lefthand plot the blue line, for $P(S = 1|Y)$,

Target	**Years**	**Lines**	**New sense**	**OED-first**	**tracks date**	**EM date**	**within** 10%
mouse	1950-2008	910k	computer pointing devince	1965	1983	1982	yes
surf	1950-2008	183k	exploring internet	1992	1994	1993	yes
boot	1920-2008	1286k	computer start up	1980	1980	1983	yes
compile	1950-2008	689k	transform to machine code	1952	1966	1966	yes
strike	1800-2008	5053k	industrial action	1810	1880	1866	yes
ostensible	1800-2008	130k	NA	–	–	none	–
present	1850-2008	56333k	NA	–	–	none	–

Table 1: The test items, their new senses and dating information – see text for explanation of columns

shows an emergent sense, departing from near zero first around 1983. The righthand plot 'tracks' plots also show a sharp increase around 1983.

We wrote an *inflection* function which applied to an inferred $P(S|Y)$ parameter returns the date, if any, at which the sense probability starts to depart from, and continues to climb from, zero and the 'EM date' column of Table 1 gives the inflection points obtained on a particular inferred $P(S|Y)$ parameter for each test item. For *mouse*, the EM-based emergence date is later than the OED first citation date, and very close to the tracks-based date. This illustrates why simply taking the OED first citation date as a gold standard would be a mistake. The OED first citation comes from a research paper in 1965, but the mouse computer peripheral only became popular considerably later and it is not unexpected that the date at which this use of the term *mouse* departed and continued to climb from zero is substantially later. For all the test items, the 'within 10%' column indicates the agreement between the EM- and tracks-based dates.

Also in Fig. 2(a-c), the box below shows for the apparently emerging sense S, the top 30 words when ranked according to the ratio of $P(w|S)$ to $P_{corp}(w)$ (call this $gist(S)$). For the *mouse* case, $gist(S = 1)$ seems mostly consistent with the 'pointing device' sense.

surf Figure 2(b). EM was run on data from 1950 to 2010, for 5 senses. In the lefthand plot the green line, for $P(S = 3|Y)$, is an emergent sense, appearing to depart from near zero first around 1993. The 'tracks' plots to the right seem to increase around around 1994. The OED first citation date of 1992 is consistent with both. The 'gist' words for $S = 3$ also seem mostly consistent the 'internet exploration' sense.

strike Figure 2(c): EM was run on data from 1800 to 2008, for 3 senses, a far longer time span than the preceding cases. In the left-hand plot, the blue line, for $P(S = 1|Y)$, is an emergent sense, appearing to depart from near zero first around 1865. The tracks plot for words intuitively associated with the industrial action sense indicate an increase from around 1880. Both of these are consistent with the OED first citation date of 1820. The 'gist' words for $S = 1$ seem consistent with $S = 1$ representing the 'industrial action' sense. For space reasons, the *boot* and *compile* outcomes are not shown in Fig.2, but analogous outcomes were obtained, summarised in Table 1.

ostensible and **present** Figure 2(d) shows the plots of the inferred $P(S|Y)$ distributions. Neither shows an emerging sense, in line with expectations.

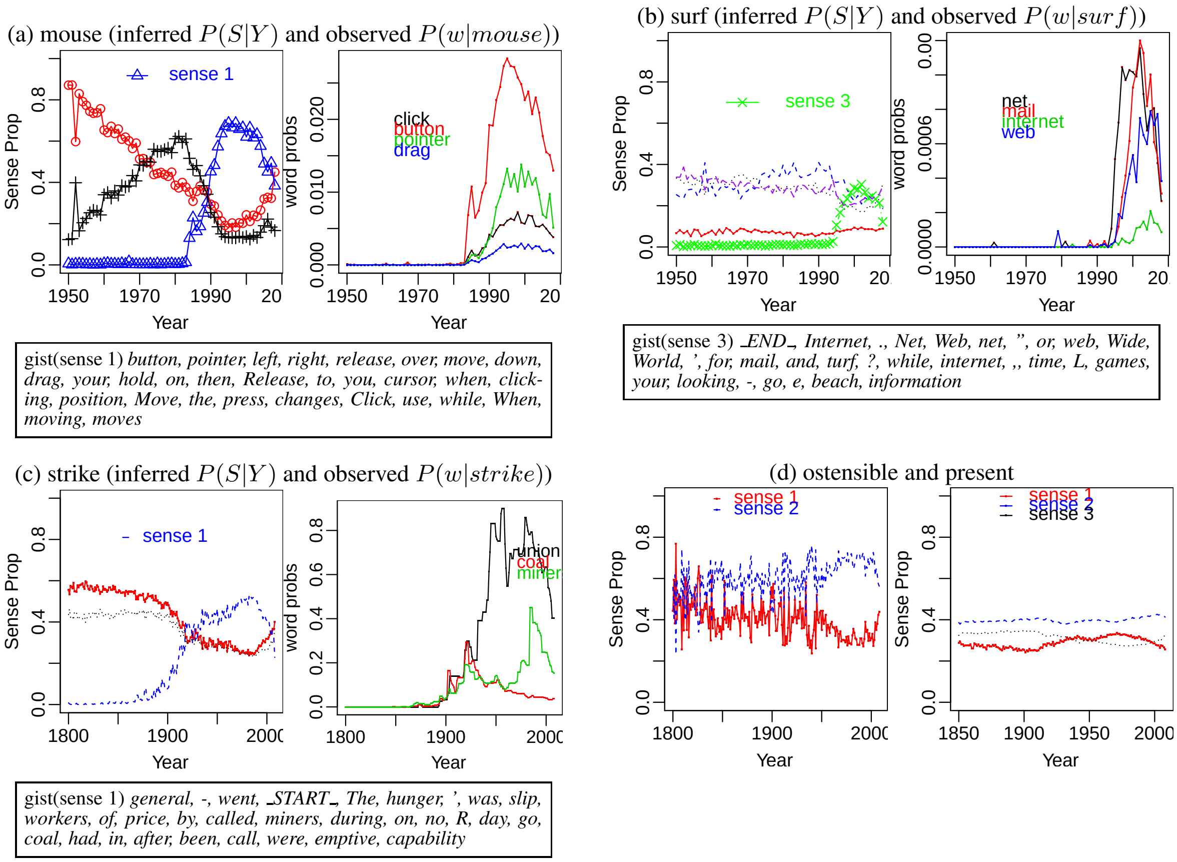
The number of senses sought in the EM runs for the different target items varied somewhat (between 3 and 5). This is somewhat to be expected as the extent of ambiguity probably varies for the different target items and in some cases where an emergence was less clear with n senses, it became clearer with $n + 1$ senses.

The procedure was implemented in C++. To obtain the code or data see `www.scss.tcd.ie/Martin.Emms/SenseDynamics`. As an indication of exection time, for a data-set of approximately 900k lines a single EM iteration takes 8.01 seconds.

4 Prior Work

In this section we review some related work on novel sense detection. Wijaya and Yeniterzi (2011) did make some analyses on the Google n-gram data in relation to indicators of sense change, seeking to apply ideas from *LDA* (Blei et al., 2003). It is not however a concrete proposal for novel sense detection and to suit the LDA perspective they artificially concatenate a year's worth of n-grams to create a single document.

In a number of papers (Lau et al., 2012; Cook et al., 2013; Cook et al., 2014) have applied a



gist(sense 1) *button, pointer, left, right, release, over, move, down, drag, your, hold, on, then, Release, to, you, cursor, when, clicking, position, Move, the, press, changes, Click, use, while, When, moving, moves*

gist(sense 3) *_END_, Internet, ., Net, Web, net, ", or, web, Wide, World, ', for, mail, and, turf, ?, while, internet, ,, time, L, games, your, looking, -, go, e, beach, information*

gist(sense 1) *general, -, went, _START_, The, hunger, ', was, slip, workers, of, price, by, called, miners, during, on, no, R, day, go, coal, had, in, after, been, call, were, emptive, capability*

Figure 2: For (a-c), lefthand plot shows EM-inferred values for $P(S|Y)$, with the sense number S of an apparent emergent sense labelled; the righthand plot show probability 'tracks' for some words intuitively associated with the emergent sense (see text for further details). $gist(S)$ in the boxes is the top-30 as ranked by $P(w|S)/P_{corp}(w)$; (d) shows inferred $P(S|Y)$ for items with no sense emergence

pre-existing word sense induction (WSI) system to novel sense detection. They draw on ideas from *LDA*, treating each context of n words as generated from n topics, and define a target's *sense* to be the most frequent *topic* amongst the context words. A striking difference to our model is that their model essentially has no parameters referring to time. They therefore take time-stamped data and pool it to train a time-unaware system, then assign target expressions their most probable sense, and finally inspect the pattern of assigned senses and the time stamps. As data, rather than a real year-by-year diachronic corpus, they worked with data representing two time *periods*, T_1 and T_2. Presumbaly in the ideal case, if a sense S truly emerged between T_1 and T_2 it should assign sense S only to items from T_2. Their approach to evaluation uses dictionary inclusion dates (ie., D^i_0 see section 2.1), so refers to senses which came to be included in dictionaries between the investigated time periods. Relative to a set of distractor expressions not thought to exhibit sense emergence their system has some success in distinguishing between true cases and the distractors in so far as when items are ranked by the ratio $p_2 : p_1$ of their inferred sense frequencies in the periods T_2 and T_1, the neologisms tend to be more highly ranked than the distractors.

Mitra et al. (2014) have also presented work on sense emergence. Their data set is a dependency-parsed version of the Google 5-grams, whose time-line they divide into eras containing equal amounts of data. They do not propose a probabilistic model but instead have a distance-based clustering system. On each era's data they perform a clustering of occurrences, and then propose ways to relate the clusters for era T_1, $\{s^{T_1}_1, \ldots s^{T_1}_m\}$ to the clusters of a later era T_2, $\{s^{T_2}_1, \ldots s^{T_2}_n\}$: if a T_2 cluster, $s^{T_2}_i$, relates to no T_1 cluster, this is counted as evidence of the emergence of a sense between T_1 and T_2. For evaluation purposes they considered the inferred emergences between the eras 1909-1953 and 2002-2005, verifying not by reference to dictionaries but by reference to the intuitions of one of the authors. Some of the examples given in the paper (eg. organ-related use of *donation*) seem plausible, others are noted as true emergences though the OED would suggest otherwise (eg. an assailant sense of *thug*).

5 Conclusions

We have proposed a relatively simple generative model, one which assumes that *given* the target's sense, the context words are independent of time – the $P(w|S)$ parameter – and that for each time, there is a prior for each sense – the $P(S|Y)$ parameter. Together this models the way the context words for a given target *do* vary over time, through the sum $\sum_S [p(S|Y)p(\vec{W}|S)]$. We have proposed an EM procedure for the estimation of the parameters of such a model and how this can be applied to Google n-gram data, which gives counts for time-stamped n-grams, and we have shown that intuitive values for the $P(S|Y)$ parameters can be obtained.

The approach is in several respects simpler than some related work discussed in section 4. To emphasize this further, the algorithm used raw rather than lemmatized words and used a data-set with no syntactic annotation. It also lacks any inherent constraint to keep $P(S|Y_1)$ and $P(S|Y_2)$ close when Y_1 and Y_2 are close and its striking that the inferred $P(S|Y)$ come out as relatively smooth functions of time. Nonetheless a direction of future work could be to consider modeling such a smoothness constraint.

Section 4 mentioned the LDA-based model of (Lau et al., 2012; Cook et al., 2013; Cook et al., 2014) and it would certainly be of interest to seek to apply their algorithms to the data we have considered and conversely, our algorithms to their data. Also the work of Mitra et al. (2014) suggests a line of development in which we adapt our approach to first estimate separate models on data belonging to eras and then similarly attempt to relate the obtained collections of word distributions.

Acknowledgments

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 993–1022,, March.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, , and Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex 2013*.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, page 1624–1635. ACL, August.

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 and Arun Jayapal. 2014. Detecting change and emergence for multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 89–93, Gothenburg, Sweden. Association for Computational Linguistics.

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.

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 2012)*, pages 591–601, Avignon, France, April.

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

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, page 1020–1029. Association for Computational Linguistics, June.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web*, DETECT '11, pages 35–40, New York, NY, USA. ACM.