A Bayesian Model of Diachronic Meaning Change

Lea Frermann and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
l.frermann@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Word meanings change over time and an automated procedure for extracting this information from text would be useful for historical exploratory studies, information retrieval or question answering. We present a dynamic Bayesian model of diachronic meaning change, which infers temporal word representations as a set of senses and their prevalence. Unlike previous work, we explicitly model language change as a smooth, gradual process. We experimentally show that this modeling decision is beneficial: our model performs competitively on meaning change detection tasks whilst inducing discernible word senses and their development over time. Application of our model to the SemEval-2015 temporal classification benchmark datasets further reveals that it performs on par with highly optimized task-specific systems.

1 Introduction

Language is a dynamic system, constantly evolving and adapting to the needs of its users and their environment (Aitchison, 2001). Words in all languages naturally exhibit a range of senses whose distribution or prevalence varies according to the genre and register of the discourse as well as its historical context. As an example, consider the word `cute` which according to the Oxford English Dictionary (OED, Stevenson 2010) first appeared in the early 18th century and originally meant *clever* or *keen-witted*.[1] By the late 19th century `cute` was used in the same sense as *cunning*. Today it mostly refers to objects or people perceived as *attractive, pretty* or *sweet*. Another example is the word `mouse` which initially was only used in the *rodent* sense. The OED dates the *computer pointing device* sense of `mouse` to 1965. The latter sense has become particularly dominant in recent decades due to the ever-increasing use of computer technology.

The arrival of large-scale collections of historic texts (Davies, 2010) and online libraries such as the Internet Archive and Google Books have greatly facilitated computational investigations of language change. The ability to automatically detect how the meaning of words evolves over time is potentially of significant value to lexicographic and linguistic research but also to real world applications. Time-specific knowledge would presumably render word meaning representations more accurate, and benefit several downstream tasks where semantic information is crucial. Examples include information retrieval and question answering, where time-related information could increase the precision of query disambiguation and document retrieval (e.g., by returning documents with newly created senses or filtering out documents with obsolete senses).

In this paper we present a dynamic Bayesian model of diachronic meaning change. Word meaning is modeled as a set of senses, which are tracked over a sequence of contiguous time intervals. We infer *temporal* meaning representations, consisting of a word's senses (as a probability distribution over words) and their relative prevalence. Our model is thus able to detect that `mouse` had one sense until the mid-20th century (characterized by words such as {cheese, tail, rat}) and subsequently acquired a

[1] Throughout this paper we denote `words` in true type, their *senses* in italics, and sense-specific context words as {lists}.

Transactions of the Association for Computational Linguistics, vol. 4, pp. 31–45, 2016. Action Editor: Tim Baldwin.
Submission batch: 12/2015; Revision batch: 2/2016; Published 2/2016.

second sense relating to computer device. Moreover, it infers subtle changes within a single sense. For instance, in the 1970s the words {cable, ball, mousepad} were typical for the *computer device* sense, whereas nowadays the terms {optical, laser, usb} are more typical. Contrary to previous work (Mitra et al., 2014; Mihalcea and Nastase, 2012; Gulordava and Baroni, 2011) where temporal representations are learnt in isolation, our model assumes that adjacent representations are co-dependent, thus capturing the nature of meaning change being fundamentally smooth and gradual (McMahon, 1994). This also serves as a form of smoothing: temporally neighboring representations influence each other if the available data is sparse.

Experimental evaluation shows that our model (a) induces temporal representations which reflect word senses and their development over time, (b) is able to detect meaning change between two time periods, and (c) is expressive enough to obtain useful features for identifying the time interval in which a piece of text was written. Overall, our results indicate that an explicit model of temporal dynamics is advantageous for tracking meaning change. Comparisons across evaluations and against a variety of related systems show that despite not being designed with any particular task in mind, our model performs competitively across the board.

2 Related Work

Most work on diachronic language change has focused on detecting whether and to what extent a word's meaning changed (e.g., between two epochs) without identifying word senses and how these vary over time. A variety of methods have been applied to the task ranging from the use of statistical tests in order to detect significant changes in the distribution of terms from two time periods (Popescu and Strapparava, 2013; Cook and Stevenson, 2010), to training distributional similarity models on time slices (Gulordava and Baroni, 2011; Sagi et al., 2009), and neural language models (Kim et al., 2014; Kulkarni et al., 2015). Other work (Mihalcea and Nastase, 2012) takes a supervised learning approach and predicts the time period to which a word belongs given its surrounding context.

Bayesian models have been previously developed for various tasks in lexical semantics (Brody and Lapata, 2009; Ó Séaghdha, 2010; Ritter et al., 2010) and word meaning change detection is no exception. Using techniques from non-parametric topic modeling, Lau et al. (2012) induce word senses (aka. topics) for a given target word over two time periods. Novel senses are then are detected based on the discrepancy between sense distributions in the two periods. Follow-up work (Cook et al., 2014; Lau et al., 2014) further explores methods for how to best measure this sense discrepancy. Rather than inferring word senses, Wijaya and Yeniterzi (2011) use a Topics-over-Time model and k-means clustering to identify the periods during which selected words move from one topic to another.

A non-Bayesian approach is put forward in Mitra et al. (2014, 2015) who adopt a graph-based framework for representing word meaning (see Tahmasebi et al. (2011) for a similar earlier proposal). In this model words correspond to nodes in a semantic network and edges are drawn between words sharing contextual features (extracted from a dependency parser). A graph is constructed for each time interval, and nodes are clustered into senses with Chinese Whispers (Biemann, 2006), a randomized graph clustering algorithm. By comparing the induced senses for each time slice and observing intercluster differences, their method can detect whether senses emerge or disappear.

Our work draws ideas from dynamic topic modeling (Blei and Lafferty, 2006b) where the evolution of topics is modeled via (smooth) changes in their associated distributions over the vocabulary. Although the dynamic component of our model is closely related to previous work in this area (Mimno et al., 2008), our model is specifically constructed for capturing sense rather than topic change. Our approach is conceptually similar to Lau et al. (2012). We also learn a joint sense representation for multiple time slices. However, in our case the number of time slices in not restricted to two and we explicitly model temporal dynamics. Like Mitra et al. (2014, 2015), we model how senses change over time. In our model, temporal representations are not *independent*, but influenced by their temporal neighbors, encouraging smooth change over time. We therefore induce a *global* and consistent set of temporal representations for each word. Our model is knowledge-lean (it does not make use of a parser) and language

independent (all that is needed is a time-stamped corpus and tools for basic pre-processing). Contrary to Mitra et al. (2014, 2015), we do not treat the tasks of inferring a semantic representation for words and their senses as two separate processes.

Evaluation of models which detect meaning change is fraught with difficulties. There is no standard set of words which have undergone meaning change or benchmark corpus which represents a variety of time intervals and genres, and is thematically consistent. Previous work has generally focused on a few hand-selected words and models were evaluated qualitatively by inspecting their output, or the extent to which they can detect meaning changes from two time periods. For example, Cook et al. (2014) manually identify 13 target words which undergo meaning change in a focus corpus with respect to a reference corpus (both news text). They then assess how their models fare at learning sense differences for these targets compared to distractors which did not undergo meaning change. They also underline the importance of using thematically comparable reference and focus corpora to avoid spurious differences in word representations.

In this work we evaluate our model's ability to detect and quantify meaning change across several time intervals (not just two). Instead of relying on a few hand-selected target words, we use larger sets sampled from our learning corpus or found to undergo meaning change in a judgment elicitation study (Gulordava and Baroni, 2011). In addition, we adopt the evaluation paradigm of Mitra et al. (2014) and validate our findings against WordNet. Finally, we apply our model to the recently established SemEval-2015 diachronic text evaluation subtasks (Popescu and Strapparava, 2015). In order to present a consistent set of experiments, we use our own corpus throughout which covers a wider range of time intervals and is compiled from a variety of genres and sources and is thus thematically coherent (see Section 4 for details). Wherever possible, we compare against prior art, with the caveat that the use of a different underlying corpus unavoidably influences the obtained semantic representations.

3 A Bayesian Model of Sense Change

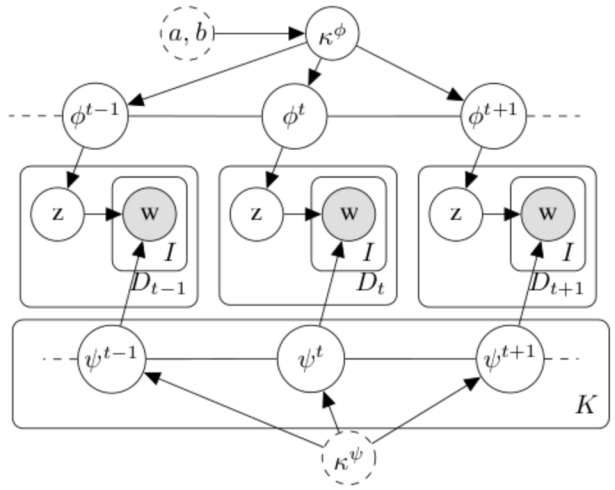
In this section we introduce SCAN, our dynamic Bayesian model of Sense ChANge. SCAN captures how a word's senses evolve over time (e.g., whether new senses emerge), whether some senses become more or less prevalent, as well as phenomena pertaining to individual senses such as meaning extension, shift, or modification. We assume that time is discrete, divided into contiguous intervals. Given a word, our model infers its senses for each time interval and their probability. It captures the gradual nature of meaning change explicitly, through dependencies between temporally adjacent meaning representations. Senses themselves are expressed as a probability distribution over words, which can also change over time.

3.1 Model Description

We create a SCAN model for each target word c. The input to the model is a corpus of short text snippets, each consisting of a mention of the target word c and its local context $\mathbf{w}$ (in our experiments this is a symmetric context window of ± 5 words). Each snippet is annotated with its year of origin. The model is parametrized with regard to the number of senses $k \in [1...K]$ of the target word c, and the length of time intervals ΔT which might be finely or coarsely defined (e.g., spanning a year or a decade).

We conflate all documents originating from the same time interval $t \in [1...T]$ and infer a temporal representation of the target word per interval. A temporal meaning representation for time t is (a) a K-dimensional multinomial distribution over word senses $\boldsymbol{\phi}^t$ and (b) a V-dimensional distribution over the vocabulary $\boldsymbol{\psi}^{t,k}$ for each word sense k. In addition, our model infers a precision parameter κ^{ϕ}, which controls the extent to which sense prevalence changes for word c over time (see Section 3.2 for details on how we model temporal dynamics).

We place individual logistic normal priors (Blei and Lafferty, 2006a) on our multinomial sense distributions $\boldsymbol{\phi}$ and sense-word distributions $\boldsymbol{\psi}^k$. A draw from the logistic normal distribution consists of (a) a draw of an n-dimensional random vector $\boldsymbol{x}$ from the multivariate normal distribution parametrized by an n-dimensional mean vector $\boldsymbol{\mu}$ and a $n \times n$ variance-covariance matrix Σ, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$; and (b) a mapping of the drawn parameters to the simplex through the logistic transformation $\phi_n = exp(x_n)/\sum_{n'} exp(x_{n'})$, which ensures a draw of valid multinomial parameters. The normal distributions are parametrized to encourage smooth



Draw $\kappa^{\phi} \sim Gamma(a, b)$
for time interval $t = 1..T$ **do**
  Draw sense distribution
  $\phi^t|\boldsymbol{\phi}^{-t}, \kappa^{\phi} \sim \mathcal{N}(\frac{1}{2}(\phi^{t-1} + \phi^{t+1}), \kappa^{\phi})$
  for sense $k = 1..K$ **do**
    Draw word distribution
    $\psi^{t,k}|\boldsymbol{\psi}^{-t}, \kappa^{\psi} \sim \mathcal{N}(\frac{1}{2}(\psi^{t-1,k} + \psi^{t+1,k}), \kappa^{\psi})$
  for document $d = 1..D$ **do**
    Draw sense $z^d \sim Mult(\boldsymbol{\phi}^t)$
    for context position $i = 1..I$ **do**
      Draw word $w^{d,i} \sim Mult(\boldsymbol{\psi}^{t,z^d})$

Figure 1: Left: plate diagram for the dynamic sense model for three time steps $\{t-1, t, t+1\}$. Constant parameters are shown as dashed nodes, latent variables as clear nodes, and observed variables as gray nodes. Right: the corresponding generative story.

change in multinomial parameters, over time (see Section 3.2 for details), and the extent of change is controlled through a precision parameter κ. We learn the value of κ^{ϕ} during inference, which allows us to model the extent of temporal change in sense prevalence individually for each target word. We draw κ^{ϕ} from a conjugate Gamma prior. We do not infer the sense-word precision parameter κ^{ψ} on all $\boldsymbol{\psi}^k$. Instead, we fix it at a high value, triggering little variation of word distributions within senses. This leads to senses being thematically coherent over time.

We now describe the generative story of our model, which is depicted in Figure 1 (right), alongside its plate diagram representation (left). First, we draw the sense precision parameter κ^{ϕ} from a Gamma prior. For each time interval t we draw (a) a multinomial distribution over senses $\boldsymbol{\phi}^t$ from a logistic normal prior; and (b) a multinomial distribution over the vocabulary $\boldsymbol{\psi}^{t,k}$ for each sense k, from another logistic normal prior. Next, we generate time-specific text snippets. For each snippet d, we first observe the time interval t, and draw a sense z^d from $Mult(\boldsymbol{\phi}^t)$. Finally, we generate I context words $w^{d,i}$ independently from $Mult(\boldsymbol{\psi}^{t,z^d})$.

3.2 Background on iGMRFs

Let $\boldsymbol{\phi} = \{\phi^1 ... \phi^T\}$ denote a T-dimensional random vector, where each ϕ^t might for example correspond to a sense probability at time t. We define a prior which encourages smooth change of parameters at neighboring times, in terms of a first order random walk on the line (graphically shown in Figure 2, and the chains of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ in Figure 1(left)). Specifically, we define this prior as an intrinsic Gaussian Markov Random Field (iGMRF; Rue and Held 2005), which allows us to model the *change* of adjacent parameters as drawn from a normal distribution, e.g.:

$$\Delta\phi^t \sim \mathcal{N}(0, \kappa^{-1}). \quad (1)$$

The iGMRF is defined with respect to the graph in Figure 2; it is sparsely connected with only first-order dependencies which allows for efficient inference. A second feature, which makes iGMRFs popular as priors in Bayesian modeling, is the fact that they can be defined purely in terms of the local changes between dependent (i.e., adjacent) variables, without the need to specify an overall mean of the model. The full conditionals explicitly capture these intuitions:

$$\phi^t|\boldsymbol{\phi}^{-t}, \kappa \sim \mathcal{N}\Big(\frac{1}{2}(\phi_{t-1} + \phi_{t+1}), \frac{1}{2\kappa}\Big), \quad (2)$$

for $1 < t < T$, where $\boldsymbol{\phi}^{-t}$ is the vector $\boldsymbol{\phi}$ except element ϕ^t and κ is a precision parameter. The value of parameter ϕ^t is distributed normally, centered around the mean of the values of its neighbors, without reference to a global mean. The precision parameter κ controls the extent of variation: how tightly coupled are the neighboring parameters? Or,

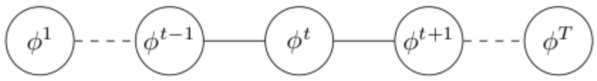


Figure 2: A linear chain iGMRF.

in our case: how tightly coupled are temporally adjacent meaning representations of a word c? We estimate the precision parameter κ^ϕ during inference. This allows us to flexibly capture sense variation over time individually for each target word.

For a detailed introduction to (i)GMRFs we refer the interested reader to Rue and Held (2005). For an application of iGMRFs to topic models see Mimno et al. (2008).

3.3 Inference

We use a blocked Gibbs sampler for approximate inference. The logistic normal prior is not conjugate to the multinomial distribution. This means that the straightforward parameter updates known for sampling standard, Dirichlet-multinomial, topic models do not apply. However, sampling-based methods for logistic normal topic models have been proposed in the literature (Mimno et al., 2008; Chen et al., 2013).

At each iteration, we sample: (a) document-sense assignments, (b) multinomial parameters from the logistic normal prior, and (c) the sense precision parameter from a Gamma prior. Our blocked sampler first iterates over all input text snippets d with context $\mathbf{w}$, and re-samples their sense assignments under the current model parameters $\{\boldsymbol{\phi}\}^T$ and $\{\boldsymbol{\psi}\}^{K\times T}$,

$$p(z^d|\mathbf{w}, t, \phi, \psi) \propto p\Big(z^d|t\Big)p\Big(\mathbf{w}|t, z^d\Big) = \phi^t_{z^d} \prod_{w\in\mathbf{w}} \psi^{t,z^d}_w \quad (3)$$

Next, we re-sample parameters $\{\boldsymbol{\phi}\}^T$ and $\{\boldsymbol{\psi}\}^{K\times T}$ from the logistic normal prior, given the current sense assignments. We use the auxiliary variable method proposed in Mimno et al. (2008) (see also Groenewald and Mokgatlhe (2005)). Intuitively, each individual parameter (e.g., sense k's prevalence at time t, ϕ^t_k) is 'shifted' within a weighted region which is bounded by the number of times sense k was observed at time t. The weights of the region are determined by the prior, in our case the normal distributions defined by the iGMRF, which ensure an influence of temporal neighbors ϕ^{t-1}_k and ϕ^{t+1}_k on the new parameter value ϕ^t_k, and smooth temporal variation as desired. The same procedure applies to each word parameter under each {time, sense} $\psi^{t,k}_w$ (see Mimno et al. 2008 for a more detailed description of the sampler). Finally, we periodically re-sample the sense precision parameter κ^ϕ from its conjugate Gamma prior.

Corpus	years covered	#words
COHA	1810–2009	142,587,656
DTE	1700–2010	124,771
CLMET3.0	1710–1810	4,531,505

Table 1: Size and coverage of our three training corpora (after pre-processing).

4 The DATE Corpus

Before presenting our evaluation we describe the corpus used as a basis for the experiments performed in this work. We applied our model to a DiAchronic TExt corpus (DATE) which collates documents spanning years 1700–2010 from three sources: (a) the COHA corpus[2] (Davies, 2010), a large collection of texts from various genres covering the years 1810–2010; (b) the training data provided by the DTE task[3] organizers (see Section 8); and (c) the portion of the CLMET3.0[4] corpus (Diller et al., 2011) corresponding to the period 1710–1810 (which is not covered by the COHA corpus and thus underrepresented in our training data). CLMET3.0 contains texts representative of a range of genres including narrative fiction, drama, letters, and was collected from various online archives. Table 1 provides details on the size of our corpus.

Documents were clustered by their year of publication as indicated in the original corpora. In the CLMET3.0 corpus, occasionally a range of years would be provided. In this case we used the final year of the range. We tokenized, lemmatized, and part of speech tagged DATE using the NLTK (Bird et al., 2009). We removed stopwords and function words. After preprocessing, we extracted target

[2] http://corpus.byu.edu/coha/

[3] http://alt.qcri.org/semeval2015/task7/index.php?id=data-and-tools

[4] http://www.kuleuven.be/~u0044428/clmet3_0.htm

word-specific input corpora for our models. These consisted of mentions of a target c and its surrounding context, a symmetric window of $\pm$ 5 words.

5 Experiment 1: Temporal Dynamics

As discussed earlier our model departs from previous approaches (e.g., Mitra et al. 2014) in that it learns globally consistent temporal representations for each word. In order to assess whether temporal dependencies are indeed beneficial, we implemented a stripped-down version of our model (SCAN-NOT) which does not have any temporal dependencies between individual time steps (i.e., without the chain iGMRF priors). Word meaning is still represented as senses and sense prevalence is modeled as a distribution over senses for each time interval. However, time intervals are now independent. Inference works as described in Section 3.3, without having to learn the κ precision parameters.

Models and Parameters We compared the two models in terms of their predictive power. We split the DATE corpus into a training period $\{d^1 ... d^t\}$ of time slices 1 through t and computed the likelihood $p(d^{t+1}|\boldsymbol{\phi}^t, \boldsymbol{\psi}^t)$ of the data at test time slice $t+1$, under the parameters inferred for the previous time slice. The time slice size was set to $\Delta T = 20$ years. We set the number of senses to $K = 8$, the word precision parameter $\kappa^{\psi} = 10$, a high value which enforces individual senses to remain thematically consistent across time. We set the initial sense precision parameter $\kappa^{\phi} = 4$, and the Gamma parameters $a = 7$ and $b = 3$. These parameters were optimized once on the development data used for the task-based evaluation discussed in Section 8. Unless otherwise specified all experiments use these values. No parameters were tuned on the test set for any task. In all experiments we ran the Gibbs sampler for 1,000 iterations, and resampled κ^{ϕ} after every 50 iterations, starting from iteration 150. We used the final state of the sampler throughout. We randomly selected 50 mid-frequency target concepts from a larger set of target concepts described in Section 8. Predictive loglikelihood scores were averaged across concepts and were calculated as the average under 10 parameter samples $\{\boldsymbol{\phi}^t, \boldsymbol{\psi}^t\}$ from the trained models.

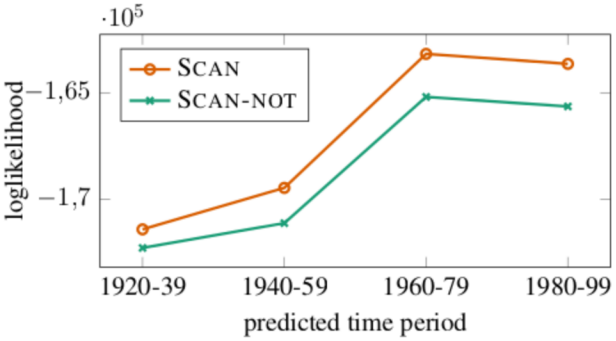


Figure 3: Predictive log likelihood of SCAN and a version without temporal dependencies (SCAN-NOT) across various test time periods.

Results Figure 3 displays predictive loglikelihood scores for four test time intervals. SCAN outperforms its stripped-down version throughout (higher is better). Since the representations learnt by SCAN are influenced (or smoothed) by neighboring representations, they overfit specific time intervals less which leads to better predictive performance. Figure 4 further shows how SCAN models meaning change for the words `band`, `power`, `transport` and `bank`. The sense distributions over time are shown as a sequence of stacked histograms, senses themselves are color-coded (and enumerated) below, in the same order as in the histograms. Each sense k is illustrated as the 10 words w assigned the highest posterior probability, marginalizing over the time-specific representations $p(w|k) = \sum_t \psi_w^{t,k}$. Words representative of prevalent senses are highlighted in bold face.

Figure 4 (top left) demonstrates that the model is able to capture various senses of the word `band`, such as *strip used for binding* (yellow bars/number 3 in the figure) or *musical band* (grey/1, orange/7). Our model predicts an increase in prevalence over the modeled time period for both senses. This is corroborated by the OED which provides the majority of references for the *binding strip* sense for the 20th century and dates the *musical band* sense to 1812. In addition a *social band* sense (violet/6, darkgreen/8; in the sense of bonding) emerges, which is present across time slices. The sense colored brown/2 refers to the *British Band*, a group of native Americans involved in the Black Hawk War in 1832, and the model indeed indicates a prevalence of this sense around this time (see bars 1800–1840 in the figure).

For the word `power` (Figure 4 (top right)),

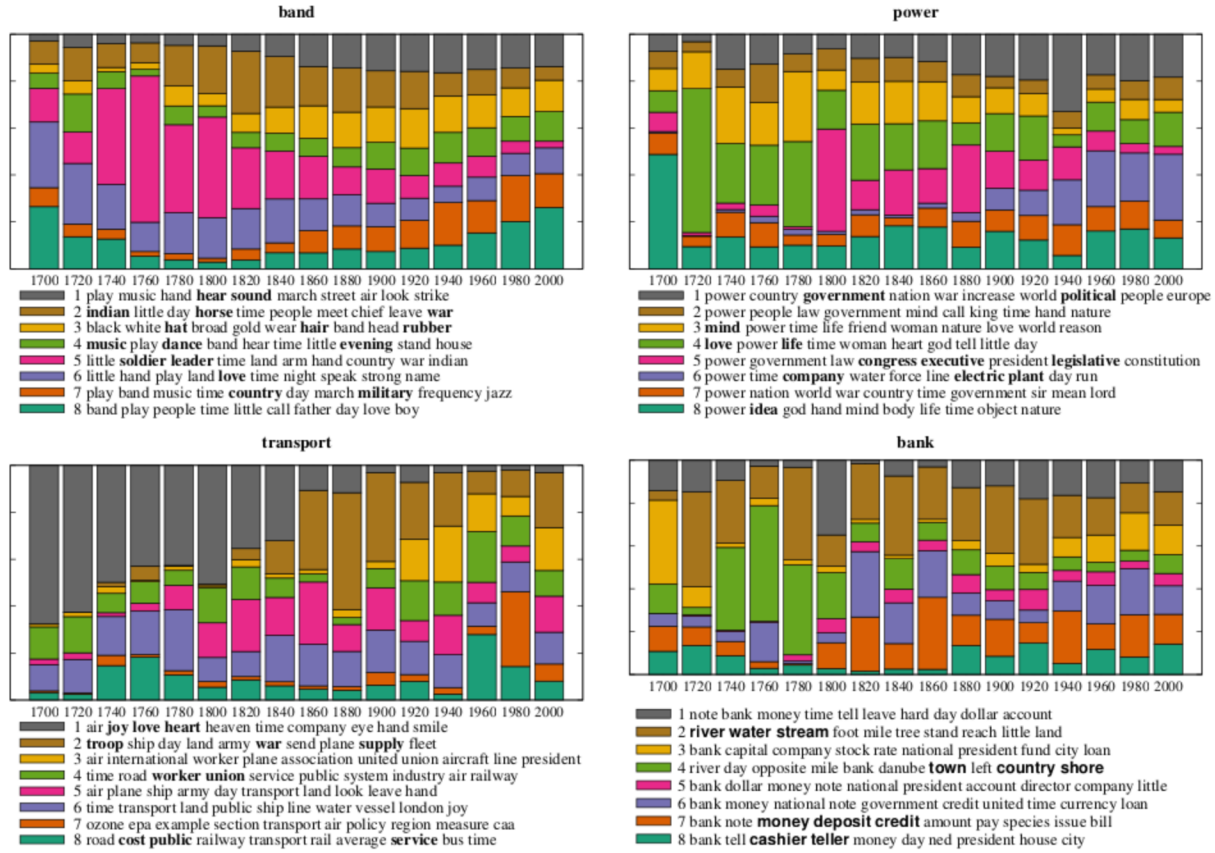


Figure 4: Tracking meaning change for the words band, power, transport and bank over 20-year time intervals between 1700 and 2010. Each bar shows the proportion of each sense (color-coded) and is labeled with the start year of the respective time interval. Senses are shown as the 10 most probable words, and particularly representative words are highlighted for illustration.

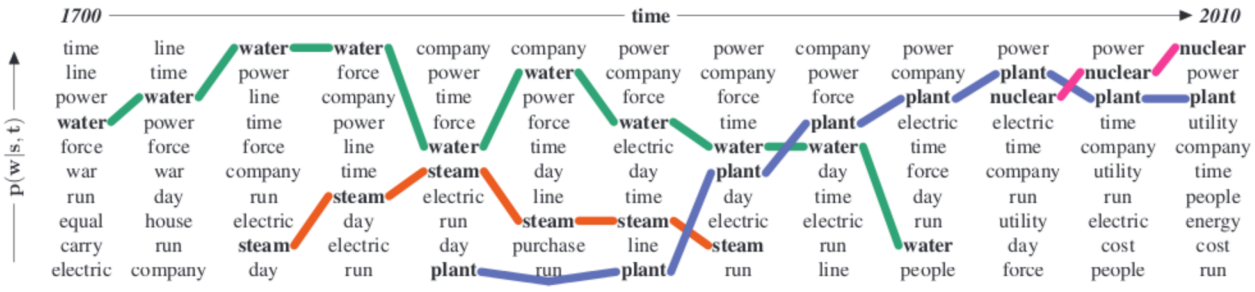


Figure 5: Sense-internal temporal dynamics for the *energy* sense of the word power (violet/6 in Figure 4). Columns show the ten most highly associated words for each time interval for the period between 1700 and 2010 (ordered by decreasing probability). We highlight how four terms characteristic of the sense develop over time (see {water, steam, plant, nuclear} in the figure).

three senses emerge: the *institutional power* (colors gray/1, brown/2, pink/5, orange/7 in the figure), *mental power* (yellow/3, lightgreen/4, darkgreen/8), and *power as supply of energy* (violet/6). The latter is an example of a "sense birth" (Mitra et al., 2014): the sense was hardly present before the mid-19th century. This is corroborated by the OED which dates the sense to 1889, whereas the OED contains references to the remaining senses for the whole modeled time period, as predicted by our model.

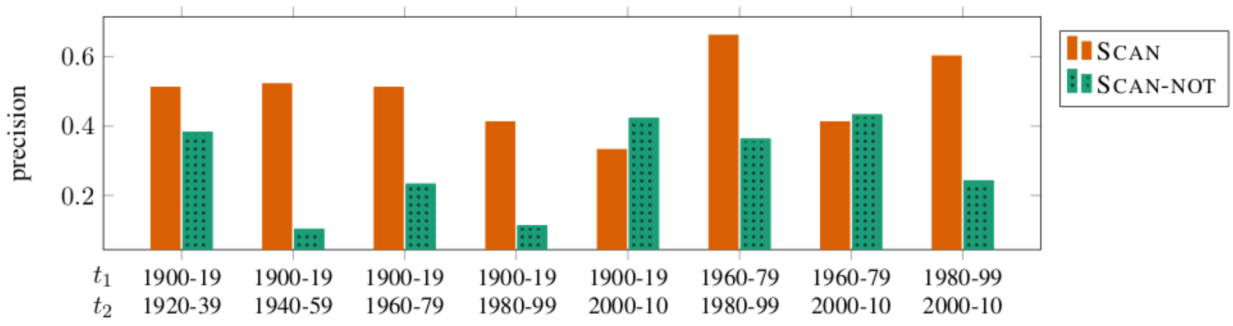


Figure 6: Precision results for the SCAN and SCAN-NOT models on the WordNet-based novel sense detection (Experiment 2). Results are shown for a selection of reference times (t_1) and focus times (t_2).

Similar trends of meaning change emerge for `transport` (Figure 4 bottom left). The bottom right plot shows the sense development for the word `bank`. Although the well-known senses *river bank* (brown/2, lightgreen/4) and *monetary institution* (rest) emerge clearly, the overall sense pattern appears comparatively stable across intervals indicating that the meaning of the word has not changed much over time.

Besides tracking sense prevalence over time, our model can also detect changes within individual senses. Because we are interested in tracking semantically stable senses, we fixed the precision parameter κ^{ψ} to a high value, to discourage too much variance within each sense. Figure 5 illustrates how the *energy* sense of the word `power` (violet/6 in Figure 4) has changed over time. Characteristic terms for a given sense are highlighted in bold face. For example, the term "water" is initially prevalent, while the term "steam" rises in prevalence towards the middle of the modeled period, and is superseded by the terms "plant" and "nuclear" towards the end.

6 Experiment 2: Novel Sense Detection

In this section and the next we will explicitly evaluate the temporal representations (i.e., probability distributions) induced by our model, and discuss its performance in the context of previous work.

Large-scale evaluation of meaning change is notoriously difficult, and many evaluations are based on limited hand-annotated goldstandard data sets. Mitra et al. (2015), however, bypass this issue by evaluating the output of their system against WordNet (Fellbaum, 1998). Here, we consider their automatic evaluation of sense-births, i.e., the emergence of novel senses. We assume that novel senses are detected at a focus time t_2 whilst being compared to a reference time t_1. WordNet is used to confirm that the proposed novel sense is indeed distinct from all other induced senses for a given word.

Method Mitra et al.'s (2015) evaluation method presupposes a system which is able to detect senses for a set of target words and identify which ones are novel. Our model does not automatically yield novelty scores for the induced senses. However, Cook et al. (2014) propose several ways to perform this task post-hoc. We use their $relevance$ score, which is based on the intuition that keywords (or collocations) which characterize the difference of a focus corpus from a reference corpus are indicative of word sense novelty.

We identify keywords for a focus corpus with respect to a reference corpus using Kilgarriff's (2009) method which is based on smoothed relative frequencies.[5] The novelty of an induced sense s can be then defined in terms of the aggregate keyword probabilities given that sense (and focus time of interest):

$$rel(s) = \sum_{w \in W} p(w|s, t_2). \quad (4)$$

where W is a keyword list and t_2 the focus time. Cook et al. (2014) suggest a straightforward extrapolation from sense novelty to word novelty:

$$rel(c) = \max_{s} rel(s), \quad (5)$$

[5]We set the smoothing parameter to $n = 10$, and like Cook et al. (2014) retrieve the top 1000 keywords.

	t_1=1900–1919 t_2=1980–1999
union	soviet united american union european war civil military people liberty
dos	system window disk pc operate program run computer de dos
entertainment	television industry program time business people world president entertainment company
station	radio station television local program network space tv broadcast air
	t_1=1960–1969 t_2=1990–1999
environmental	supra note law protection id agency impact policy factor federal
users	computer window information software system wireless drive web building available
virtual	reality virtual computer center experience week community separation increase
disk	hard disk drive program computer file store ram business embolden

Table 2: Example target terms (left) with novel senses (right) as identified by SCAN in focus corpus t_2 (when compared against reference corpus t_1). Top: terms used in novel sense detection study (Experiment 2). Bottom: terms from the Gulordava and Baroni (2011) gold standard (Experiment 3).

where $rel(c)$ is the highest novelty score assigned to any of the target word's senses. A high $rel(c)$ score suggests that a word has undergone meaning change.

We obtained candidate terms and their associated novel senses from the DATE corpus, using the $relevance$ metric described above. The novel senses from the focus period and all senses induced for the reference period, except for the one corresponding to the novel sense, were passed on to Mitra et al.'s (2015) WordNet-based evaluator which proceeds as follows. Firstly, each induced sense s is mapped to the WordNet synset u with the maximum overlap:

$$synset(s) = \arg\max_{u} \ overlap(s, u). \quad (6)$$

Next, a predicted novel sense n is deemed truly novel if its mapped synset is distinct from any synset mapped to a different induced sense:

$$\forall_{s'} synset(s') \neq synset(n). \quad (7)$$

Finally, overall precision is calculated as the fraction of sense-births confirmed by WordNet over all birth-candidates proposed by the model. Like Mitra et al. (2015) we only report results on target words for which all induced senses could be successfully mapped to a synset.

Models and Parameters We obtained the broad set of target words used for the task-based evaluation (in Section 8) and trained models on the DATE corpus. We set the number of senses $K = 4$ following Mitra et al. (2015) who note that the WordNet mapper works best for words with a small number of senses, and the time intervals to $\Delta T = 20$ as in Experiment 1. We identified 200 words[6] with highest novelty score (Equation (5)) as sense birth candidates. We compared the performance of the full SCAN model against SCAN-NOT which learns senses independently for time intervals. We trained both models on the same data with identical parameters. For SCAN-NOT, we must post-hoc identify corresponding senses across time intervals. We used the Jensen-Shannon divergence between the reference- and focus-time specific word distributions $JS(p(w|s, t_1)||p(w|s, t_2))$ and assigned each focus-time sense to the sense with smallest divergence at reference time.

Results Figure 6 shows the performance of our models on the task of sense birth detection. SCAN performs better than SCAN-NOT, underscoring the importance of joint modeling of senses across time slices and incorporation of temporal dynamics. Our accuracy scores are in the same ballpark as Mitra et al. (2014, 2015). Note, however that the scores are not directly comparable due to differences in training corpora, focus and reference times, and candidate words. Mitra et al. (2015) use the larger Google syntactic n-gram corpus, as well as richer linguistic information in terms of syntactic dependencies. We show that our model which does not rely on syntactic annotations performs competitively even when trained on smaller data. Table 2 (top) displays examples of words assigned highest novelty scores for the reference period 1900–1919 and focus period 1980–1999.

[6]This threshold was tuned on one reference-focus time pair.

7 Experiment 3: Word Meaning Change

In this experiment we evaluate whether model induced temporal word representations capture perceived word novelty. Specifically, we adopt the evaluation framework (and dataset) introduced in Gulordava and Baroni (2011)[7] and discussed below.

Method Gulordava and Baroni (2011) do not model word senses directly; instead they obtain distributional representations of words from the Google Books (bigram) data for two time slices, namely the 1960s (reference corpus) and 1990s (focus corpus). To detect change in meaning, they measure cosine similarity between the vector representations of a target word in the reference and focus corpus. It is assumed that low similarity indicates significant meaning change. To evaluate the output of their system, they created a test set of 100 target words (nouns, verbs, and adjectives), and asked five annotators to rate each word with respect to its degree of meaning change between the 1960s and the 1990s. The annotators used a 4-point ordinal scale (0: no change, 1: almost no change, 2: somewhat change, 3: changed significantly). Words were subsequently ranked according to the mean rating given by the annotators. Inter-annotator agreement on the novel sense detection task was 0.51 (pairwise Pearson correlation) and can be regarded as an upper bound on model performance.

Models and Parameters We trained models for all words in Gulordava and Baroni's (2011) gold-standard. We used the DATE subcorpus covering years 1960 through 1999 partitioned by decade ($\Delta T = 10$). The first and last time interval were defined as reference and focus time, respectively (t_1=1960–1969, t_2=1990–1999). As in Experiment 2, a novelty score was assigned to each target word (using Equation (5)). We computed Spearman's ρ rank correlations between gold standard and model rankings (Gulordava and Baroni, 2011). We trained SCAN models setting the number of senses to $K = 8$. We also trained SCAN-NOT models with identical parameters. We report results averaged over five independent parameter estimates. Finally, as in Gulordava and Baroni (2011) we compare against a frequency baseline which ranks words by their log relative frequency in the reference and focus corpus.

[7]We thank Kristina Gulordava for sharing their evaluation data set of target words and human judgments.

system	corpus	Spearman's ρ
Gulordava (2011)	Google	0.386
SCAN	DATE	0.377
SCAN-NOT	DATE	0.255
frequency baseline	DATE	0.325

Table 3: Spearman's ρ rank correlations between system novelty rankings and the human-produced ratings. All correlations are statistically significant ($p < 0.02$). Results for SCAN and SCAN-NOT are averages over five trained models.

Results The results of this evaluation are shown in Table 3. As can be seen, SCAN outperforms SCAN-NOT and the frequency baseline. For reference, we also report the correlation coefficient obtained in Gulordava and Baroni (2011) but emphasize that the scores are not directly comparable due to differences in training data: Gulordava and Baroni (2011) use the Google bigrams corpus (which is much larger compared to DATE). Table 2 (bottom) displays examples of words which achieved highest novelty scores in this evaluation, and their associated novel senses.

8 Experiment 4: Task-based Evaluation

In the previous sections we demonstrated how SCAN captures meaning change between two periods. In this section, we assess our model on an extrinsic task which relies on meaning representations spanning several time slices. We quantitatively evaluate our model on the SemEval-2015 benchmark datasets released as part of the Diachronic Text Evaluation exercise (Popescu and Strapparava 2015; DTE). In the following we first present the DTE subtasks, and then move on to describe our training data, parameter settings, and systems used for comparison to our model.

SemEval DTE Tasks Diachronic text evaluation is an umbrella term used by the SemEval-2015 organizers to represent three subtasks aiming to assess the performance of computational methods used to identify when a piece of text was written. A similar problem is tackled in Chambers (2012) who label documents with time stamps whilst focusing on

explicit time expressions and their discriminatory power. The SemEval data consists of news snippets, which range between a few words and multiple sentences. A set of training snippets, as well as gold-annotated development and test datasets are provided. DTE subtasks 1 and 2 involve temporal classification: given a news snippet and a set of non-overlapping time intervals covering the period 1700 through 2010, the system's task is to select the interval corresponding to the snippet's year of origin. Temporal intervals are consecutive and constructed such that the correct interval is centered around the actual year of origin. For both tasks temporal intervals are created at three levels of granularity (fine, medium, and coarse).

Subtask 1 involves snippets which contain an explicit cue for time of origin. The presence of a temporal cue was determined by the organizers by checking the entities' informativeness in external resources. Consider the example below:

(8) President de Gaulle favors an independent European nuclear striking force [...]

The mentions of French president de Gaulle and nuclear warfare suggest that the snippet was written after the mid-1950s and indeed it was published in 1962. A hypothetical system would then have to decide amongst the following classes:

{1700–1702, 1703–1705, ..., 1961–1963, ..., 2012–2014}
{1699–1706, 1707–1713, ..., 1959–1965, ..., 2008–2014}
{1696–1708, 1709–1721, ..., 1956–1968, ..., 2008–2020}

The first set of classes correspond to fine-grained intervals of 2-years, the second set to medium-grained intervals of 6-years and the third set to coarse-grained intervals of 12-years. For the snippet in example (8) classes 1961–1963, 1959–1965, and 1956–1968 are the correct ones.

Subtask 2 involves temporal classification of snippets which lack explicit temporal cues, but contain implicit ones, e.g., as indicated by lexical choice or spelling. The snippet in example (9) was published in 1891 and the spelling of *to-day*, which was common up to the early 20th century, is an implicit cue:

(9) The local wheat market was not quite so strong to-day as yesterday.

Analogously to subtask 1, systems must select the right temporal interval from a set of contiguous time intervals of differing granularity. For this task, which is admittedly harder, levels of temporal granularity are coarser corresponding to 6-year, 12-year and 20-year intervals.

Participating SemEval Systems We compared our model against three other systems which participated in the SemEval task.[8] AMBRA (Zampieri et al., 2015) adopts a learning-to-rank modeling approach and uses several stylistic, grammatical, and lexical features. IXA (Salaberri et al., 2015) uses a combination of approaches to determine the period of time in which a piece of news was written. This involves searching for specific mentions of time within the text, searching for named entities present in the text and then establishing their reference time by linking these to Wikipedia, using Google n-grams, and linguistic features indicative of language change. Finally, UCD (Szymanski and Lynch, 2015) employs SVMs for classification using a variety of informative features (e.g., POS-tag n-grams, syntactic phrases), which were optimized for the task through automatic feature selection.

Models and Parameters We trained our model for individual words and obtained representations of their meaning for different points in time. Our set of target words consisted of all nouns which occurred in the development datasets for DTE subtasks 1 and 2 as well as all verbs which occurred at least twice in this dataset. After removing infrequent words we were left with 883 words (out of 1,116) which we used in this evaluation. Target words were not optimized with respect to the test data in any way; it is thus reasonable to expect better performance with an adjusted set of words.

We set the model time interval to $\Delta T = 5$ years and the number of senses per word to $K = 8$. We also evaluated SCAN-NOT, the stripped-down version of SCAN, with identical parameters. Both SCAN and SCAN-NOT predict the time of origin for a test snippet as follows. We first detect mentions of target words in the snippet. Then, for each mention c we construct a document, akin to the training documents, consisting of c and its context $\mathbf{w}$, the ± 5 words surrounding c. Given $\{c, \mathbf{w}\}$, we approximate

[8] We do not report results for the system USAAR which achieved close to 100% accuracy by searching for the test snippets on the web, without performing any temporal inference.

	Task 1						*Task 2*					
	2 yr		6 yr		12 yr		6 yr		12 yr		20 yr	
	acc	*p*	*acc*	*p*	*acc*	*p*	*acc*	*p*	*acc*	*p*	*acc*	*p*
Baseline	.097	.010	.214	.017	.383	.046	.199	.025	.343	.047	.499	.057
SCAN-NOT	.265	**.086**	.435	**.139**	.609	.169	.259	.041	.403	.056	.567	.098
SCAN	**.353**	.049	**.569**	.112	**.748**	**.206**	.376	.053	.572	.091	.719	.135
IXA	.187	.020	.375	.041	.557	.090	.261	.037	.428	.067	.622	.098
AMBRA	.167	.037	.367	.071	.554	.074	.605	.143	.767	.143	.868	.292
UCD	–	–	–	–	–	–	**.759**	.463	**.846**	.472	**.910**	.542
SVM SCAN	.192	.034	.417	.097	.545	.127	.573	.331	.667	.368	.790	.428
SVM SCAN+ngram	.222	.030	.467	.079	.627	.142	.747	**.481**	.821	**.500**	.897	**.569**

Table 4: Results on Diachronic Text Evaluation Tasks 1 and 2 for a random baseline, our SCAN model, its stripped-down version without iGMRFs (SCAN-NOT), the SemEval submissions (IXA, AMBRA and UCD), and SVMs trained with SCAN features (SVM SCAN), and with additional character n-gram features (SVM SCAN+ngram). Results are shown for three levels of granularity, a strict precision measure p, and a distance-discounting measure acc.

a distribution over time intervals as:

$$p^{(c)}(t|\mathbf{w}) \propto p^{(c)}(\mathbf{w}|t) \times p^{(c)}(t) \tag{10}$$

where the superscript (c) indicates parameters from the word-specific model, we marginalize over senses and assume a uniform distribution over time slices $p^{(c)}(t)$. Finally, we combine the word-wise predictions into a final distribution $p(t) = \prod_c p^{(c)}(t|, \mathbf{w})$, and predict the time t with highest probability.

Supervised Classification We also apply our model in a supervised setting, i.e., by extracting features for classifier prediction. Specifically, we trained a multiclass SVM (Chang and Lin, 2011) on the training data provided by the SemEval organizers (for DTE tasks 1 and 2). For each observed word within each snippet, we added as feature its most likely sense k given t, the true time of origin:

$$\arg\max_k \; p^{(c)}(k|t). \tag{11}$$

We also trained a multiclass SVM which uses character n-gram ($n \in \{1, 2, 3\}$) features in addition to the model features. Szymanski and Lynch (2015) identified character n-grams as the most predictive feature for temporal text classification using SVMs. Their system (UCD) achieved the best published scores in DTE subtask 2. Following their approach, we included all n-grams that were observed more than 20 times in the DTE training data.

Results We employed two evaluation measures proposed by the DTE organizers. These are precision p, i.e., the percentage of times a system has predicted the correct time period. And accuracy acc which is more lenient, and penalizes system predictions proportional to their distance from the true interval. We compute the p and acc scores for our models using the evaluation script provided by the SemEval organizers. Table 4 summarizes our results for DTE subtasks 1 and 2. We compare SCAN against a baseline which selects a time interval at random[9] averaged over five runs. We also show results for a stripped-down version of our model without the iGMRFs (SCAN-NOT) and for the systems which participated in SemEval.

For subtask 1, the two versions of SCAN outperform all SemEval systems across the board. SCAN-NOT occasionally outperforms SCAN in the strict precision metric, however, the full SCAN model consistently achieves better accuracy scores which are more representative since they factor in the proximity of the prediction to the true value. In subtask 2, the UCD and SVM SCAN+ngram systems perform comparably. They both use SVMs for the classification task, however our own model employs a less expressive feature set based on SCAN and character n-grams, and does not take advantage of feature selection which would presumably enhance performance. With the exception of AMBRA, all other participating systems used external resources (such as Wikipedia and Google n-grams); it is thus fair to assume they had access to at least as much training data as our SCAN model. Consequently, the

[9]We recomputed the baseline scores for subtasks 1 and 2 due to inconsistencies in the results provided by the DTE organizers.

gap in performance can not solely be attributed to a difference in the size of the training data.

We also observe that IXA and SCAN, given identical granularity, perform better on subtask 1, while AMBRA and our own SVM-based systems exhibit the opposite trend. The IXA system uses a combination of knowledge sources in order to determine when a piece of news was written, including explicit mentions of temporal expressions within the text, named entities, and linked information to those named entities from Wikipedia. AMBRA on the other hand exploits more shallow stylistic, grammatical and lexical features within the learning-to-rank paradigm. An interesting direction for future work would be to investigate which features are most appropriate for different DTE tasks. Overall, it is encouraging to see that the generic temporal word representations inferred by SCAN lead to competitively performing models on both temporal classification tasks without any explicit tuning.

9 Conclusion

In this paper we introduced SCAN, a dynamic Bayesian model of diachronic meaning change. Our model learns a coherent set of co-dependent time-specific senses for individual words and their prevalence. Evaluation of the model's output showed that the learnt representations reflect (a) different senses of ambiguous words (b) different kinds of meaning change (such as new senses being established), and (c) connotational changes within senses. SCAN departs from previous work in that it models temporal dynamics explicitly. We demonstrated that this feature yields more general semantic representations as indicated by predictive loglikelihood and a variety of extrinsic evaluations. We also experimentally evaluated SCAN on novel sense detection and the SemEval DTE task, where it performed on par with the best published results, without any extensive feature engineering or task specific tuning.

We conclude by discussing limitations of our model and directions for future work. In our experiments we fix the number of senses K for all words across all time periods. Although this approach did not harm performance (even in case of SemEval where we handled more than 800 target concepts), it is at odds with the fact that words vary in their degree of ambiguity, and that word senses continuously appear and disappear. A non-parametric version of our model would infer an appropriate number of senses from the data, individually for each time period. Also note that in our experiments we used context as a bag of words. It would be interesting to explore more systematically how different kinds of contexts (e.g., named entities, multiword expressions, verbs vs. nouns) influence the representations the model learns. Furthermore, while SCAN captures the temporal dynamics of word senses, it cannot do so for words themselves. Put differently, the model cannot identify whether a new word is used which did not exist before, or that a word ceased to exist after a specific point in time. A model internal way of detecting word (dis)appearance would be desirable, especially since new terms are continuously being introduced thanks to popular culture and various new media sources.

In the future, we would like to apply our model to different text genres and levels of temporal granularity. For example, we could work with Twitter data, an increasingly popular source for opinion tracking, and use our model to identify short-term changes in word meanings or connotations.

Acknowledgments

We are grateful to the anonymous reviewers whose feedback helped to substantially improve the present paper. We thank Charles Sutton and Iain Murray for helpful discussions, and acknowledge the support of EPSRC through project grant EP/I037415/1.

References

Aitchison, Jean. 2001. *Language Change: Progress Or Decay?*. Cambridge Approaches to Linguistics. Cambridge University Press.

Biemann, Chris. 2006. Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In *Proceedings of TextGraphs: the 1st Workshop on Graph Based Methods for Natural Language Processing*. New York City, NY, USA, pages 73–80.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Blei, David M. and John D. Lafferty. 2006a. Cor-

related Topic Models. In *Advances in Neural Information Processing Systems 18*, Vancouver, BC, Canada, pages 147–154.

Blei, David M. and John D. Lafferty. 2006b. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, PA, USA, pages 113–120.

Brody, Samuel and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL*. Athens, Greece, pages 103–111.

Chambers, Nathanael. 2012. Labeling Documents with Timestamps: Learning from their Time Expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea, pages 98–106.

Chang, Chih-Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, Jianfei, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. 2013. Scalable Inference for Logistic-Normal Topic Models. In *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, pages 2445–2453.

Cook, Paul, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel Word-sense Identification. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, pages 1624–1635.

Cook, Paul and Suzanne Stevenson. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta, pages 28–34.

Davies, Mark. 2010. The Corpus of Historical American English: 400 million words, 1810-2009. Available online at `http://corpus.byu.edu/coha/`.

Diller, Hans-Jürgen, Hendrik de Smet, and Jukka Tyrkkö. 2011. A European database of descriptors of english electronic texts. *The European English messenger* 19(2):29–35.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Groenewald, Pieter C. N. and Lucky Mokgatlhe. 2005. Bayesian Computation for Logistic Regression. *Computational Statistics & Data Analysis* 48(4):857–868.

Gulordava, Kristina and Marco Baroni. 2011. A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics*. Edinburgh, Scotland, pages 67–71.

Kilgarriff, Adam. 2009. Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA, pages 61–65.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*. Geneva, Switzerland, pages 625–635.

Lau, Han Jey, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, USA, pages 259–270.

Lau, Jey Han, 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*. Avignon, France, pages 591–601.

McMahon, April M.S. 1994. *Understanding Language Change*. Cambridge University Press.

Mihalcea, Rada and Vivi Nastase. 2012. Word Epoch Disambiguation: Finding How Words Change over Time. In *Proceedings of the 50th*

Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, pages 259–263.

Mimno, David, Hanna Wallach, and Andrew McCallum. 2008. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. In *NIPS Workshop on Analyzing Graphs*. Vancouver, Canada.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21:773–798.

Mitra, Sunny, 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*. Baltimore, MD, USA, pages 1020–1029.

Ó Séaghdha, Diarmuid. 2010. Latent Variable Models of Selectional Preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pages 435–444.

Popescu, Octavian and Carlo Strapparava. 2013. Behind the Times: Detecting Epoch Changes using Large Corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan, pages 347–355.

Popescu, Octavian and Carlo Strapparava. 2015. SemEval 2015, Task 7: Diachronic Text Evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, CO, USA, pages 869–877.

Ritter, Alan, Mausam, and Oren Etzioni. 2010. A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pages 424–434.

Rue, Håvard and Leonhard Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Athens, Greece, pages 104–111.

Salaberri, Haritz, Iker Salaberri, Olatz Arregi, and Beñat Zapirain. 2015. IXAGroupEHUDiac: A Multiple Approach System towards the Diachronic Evaluation of Texts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, CO, USA, pages 840–845.

Stevenson, Angus, editor. 2010. *The Oxford English Dictionary*. Oxford University Press, third edition.

Szymanski, Terrence and Gerard Lynch. 2015. UCD: Diachronic Text Classification with Character, Word, and Syntactic N-grams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, CO, USA, pages 879–883.

Tahmasebi, Nina, Thomas Risse, and Stefan Dietze. 2011. Towards automatic language evolution tracking, A study on word sense tracking. In *Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011)*. Bonn, Germany.

Wijaya, Derry Tanti 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*. Glasgow, Scotland, UK, pages 35–40.

Zampieri, Marcos, Alina Maria Ciobanu, Vlad Niculae, and Liviu P. Dinu. 2015. AMBRA: A Ranking Approach to Temporal Text Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, CO, USA, pages 851–855.