Word Sense Induction for Novel Sense Detection

Jey Han Lau,[♠♡] **Paul Cook,**[♡] **Diana McCarthy,**[♣]
David Newman,[♢] and **Timothy Baldwin**[♠♡]
♠ NICTA Victoria Research Laboratory
♡ Dept of Computer Science and Software Engineering, University of Melbourne
♢ Dept of Computer Science, University of California Irvine
♣ Lexical Computing
jhlau@csse.unimelb.edu.au, paulcook@unimelb.edu.au,
diana@dianamccarthy.co.uk, newman@uci.edu, tb@ldwin.net

Abstract

We apply topic modelling to automatically induce *word senses* of a target word, and demonstrate that our word sense induction method can be used to automatically detect words with emergent novel senses, as well as token occurrences of those senses. We start by exploring the utility of standard topic models for word sense induction (WSI), with a pre-determined number of topics (=senses). We next demonstrate that a non-parametric formulation that learns an appropriate number of senses per word actually performs better at the WSI task. We go on to establish state-of-the-art results over two WSI datasets, and apply the proposed model to a novel sense detection task.

1 Introduction

Word sense induction (WSI) is the task of automatically inducing the different senses of a given word, generally in the form of an unsupervised learning task with senses represented as clusters of token instances. It contrasts with word sense disambiguation (WSD), where a fixed sense inventory is assumed to exist, and token instances of a given word are disambiguated relative to the sense inventory. While WSI is intuitively appealing as a task, there have been no real examples of WSI being successfully deployed in end-user applications, other than work by Schutze (1998) and Navigli and Crisafulli (2010) in an information retrieval context. A key contribution of this paper is the successful application of WSI to the lexicographical task of novel sense detection, i.e. identifying words which have taken on new senses over time.

One of the key challenges in WSI is learning the appropriate sense granularity for a given word, i.e. the number of senses that best captures the token occurrences of that word. Building on the work of Brody and Lapata (2009) and others, we approach WSI via topic modelling — using Latent Dirichlet Allocation (LDA: Blei et al. (2003)) and derivative approaches — and use the topic model to determine the appropriate sense granularity. Topic modelling is an unsupervised approach to jointly learn topics — in the form of multinomial probability distributions over words — and per-document topic assignments — in the form of multinomial probability distributions over topics. LDA is appealing for WSI as it both assigns senses to words (in the form of topic allocation), and outputs a representation of each sense as a weighted list of words. LDA offers a solution to the question of sense granularity determination via non-parametric formulations, such as a Hierarchical Dirichlet Process (HDP: Teh et al. (2006), Yao and Durme (2011)).

Our contributions in this paper are as follows. We first establish the effectiveness of HDP for WSI over both the SemEval-2007 and SemEval-2010 WSI datasets (Agirre and Soroa, 2007; Manandhar et al., 2010), and show that the non-parametric formulation is superior to a standard LDA formulation with oracle determination of sense granularity for a given word. We next demonstrate that our interpretation of HDP-based WSI is superior to other topic model-based approaches to WSI, and indeed, better than the best-published results for both SemEval datasets. Finally, we apply our method to the novel sense detection task based on a dataset developed in this research, and achieve highly encouraging results.

2 Methodology

In topic modelling, documents are assumed to exhibit multiple topics, with each document having

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 591–601,
Avignon, France, April 23 - 27 2012.

its own distribution over topics. Words are generated in each document by first sampling a topic from the document's topic distribution, then sampling a word from that topic. In this work we use the topic models's probabilistic assignment of topics to words for the WSI task.

2.1 Data Representation and Pre-processing

In the context of WSI, topics form our sense representation, and words in a sentence are generated conditioned on a particular sense of the target word. The "document" in the WSI case is a single sentence or a short document fragment containing the target word, as we would not expect to be able to generate a full document from the sense of a single target word.[1] In the case of the SemEval datasets, we use the word contexts provided in the dataset, while in our novel sense detection experiments, we use a context window of three sentences, one sentence to either side of the token occurrence of the target word.

As our baseline representation, we use a bag of words, where word frequency is kept but not word order. All words are lemmatised, and stopwords and low frequency terms are removed.

We also experiment with the addition of positional context word information, as commonly used in WSI. That is, we introduce an additional word feature for each of the three words to the left and right of the target word.

Padó and Lapata (2007) demonstrated the importance of syntactic dependency relations in the construction of semantic space models, e.g. for WSD. Based on these findings, we include dependency relations as additional features in our topic models,[2] but just for dependency relations that involve the target word.

2.2 Topic Modelling

Topic models learn a probability distribution over topics for each document, by simply aggregating the distributions over topics for each word in the document. In WSI terms, we take this distribution over topics for each target word ("instance" in WSI parlance) as our distribution over senses for that word.

In our initial experiments, we use LDA topic modelling, which requires us to set T, the number of topics to be learned by the model. The LDA generative process is: (1) draw a latent topic z from a document-specific topic distribution $P(t = z|d)$ then; (2) draw a word w from the chosen topic $P(w|t = z)$. Thus, the probability of producing a single copy of word w given a document d is given by:

$$P(w|d) = \sum_{z=1}^{T} P(w|t = z)P(t = z|d).$$

In standard LDA, the user needs to specify the number of topics T. In non-parametric variants of LDA, the model dynamically learns the number of topics as part of the topic modelling. The particular implementation of non-parametric topic model we experiment with is Hierarchical Dirichlet Process (HDP: Teh et al. (2006)),[3] where, for each document, a distribution of mixture components $P(t|d)$ is sampled from a base distribution G_0 as follows: (1) choose a base distribution $G_0 \sim DP(\gamma, H)$; (2) for each document d, generate distribution $P(t|d) \sim DP(\alpha_0, G_0)$; (3) draw a latent topic z from the document's mixture component distribution $P(t|d)$, in the same manner as for LDA; and (4) draw a word w from the chosen topic $P(w|t = z)$.[4]

For both LDA and HDP, we individually topic model each target word, and determine the sense assignment z for a given instance by aggregating over the topic assignments for each word in the instance and selecting the sense with the highest aggregated probability, $\arg\max_z P(t = z|d)$.

3 SemEval Experiments

To facilitate comparison of our proposed method for WSI with previous approaches, we use the dataset from the SemEval-2007 and SemEval-2010 word sense induction tasks (Agirre and

[1] Notwithstanding the one sense per discourse heuristic (Gale et al., 1992).

[2] We use the Stanford Parser to do part of speech tagging and to extract the dependency relations (Klein and Manning, 2003; De Marneffe et al., 2006).

[3] We use the C++ implementation of HDP (`http://www.cs.princeton.edu/~blei/topicmodeling.html`) in our experiments.

[4] The two HDP parameters γ and α_0 control the variability of senses in the documents. In particular, γ controls the degree of sharing of topics across documents — a high γ value leads to more topics, as topics for different documents are more dissimilar. α_0, on the other hand, controls the degree of mixing of topics within a document — a high α_0 generates fewer topics, as topics are less homogeneous within a document.

Soroa, 2007; Manandhar et al., 2010). We first experiment with the SemEval-2010 dataset, as it includes explicit training and test data for each target word and utilises a more robust evaluation methodology. We then return to experiment with the SemEval-2007 dataset, for comparison purposes with other published results for topic modelling approaches to WSI.

3.1 SemEval-2010

3.1.1 Dataset and Methodology

Our primary WSI evaluation is based on the dataset provided by the SemEval-2010 WSI shared task (Manandhar et al., 2010). The dataset contains 100 target words: 50 nouns and 50 verbs. For each target word, a fixed set of training and test instances are supplied, typically 1 to 3 sentences in length, each containing the target word.

The default approach to evaluation for the SemEval-2010 WSI task is in the form of WSD over the test data, based on the senses that have been automatically induced from the training data. Because the induced senses will likely vary in number and nature between systems, the WSD evaluation has to incorporate a sense alignment step, which it performs by splitting the test instances into two sets: a mapping set and an evaluation set. The optimal mapping from induced senses to gold-standard senses is learned from the mapping set, and the resulting sense alignment is used to map the predictions of the WSI system to pre-defined senses for the evaluation set. The particular split we use to calculate WSD effectiveness in this paper is 80%/20% (mapping/test), averaged across 5 random splits.[5]

The SemEval-2010 training data consists of approximately 163K training instances for the 100 target words, all taken from the web. The test data is approximately 9K instances taken from a variety of news sources. Following the standard approach used by the participating systems in the SemEval-2010 task, we induce senses only from the training instances, and use the learned model to assign senses to the test instances.

[5] A 60%/40% split is also provided as part of the task setup, but the results are almost identical to those for the 80%/20% split, and so are omitted from this paper. The original task also made use of V-measure and Paired F-score to evaluate the induced word sense clusters, but have degenerate behaviour in correlating strongly with the number of senses induced by the method (Manandhar et al., 2010), and are hence omitted from this paper.

In our original experiments with LDA, we set the number of topics (T) for each target word to the number of senses represented in the test data for that word (varying T for each target word). This is based on the unreasonable assumption that we will have access to gold-standard information on sense granularity for each target word, and is done to establish an upper bound score for LDA. We then relax the assumption, and use a fixed T setting for each of sets of nouns ($T = 7$) and verbs ($T = 3$), based on the average number of senses from the test data in each case. Finally, we introduce positional context features for LDA, once again using the fixed T values for nouns and verbs.

We next apply HDP to the WSI task, using positional features, but learning the number of senses automatically for each target word via the model. Finally, we experiment with adding dependency features to the model.

To summarise, we provide results for the following models:

1. **LDA+Variable** T: LDA with variable T for each target word based on the number of gold-standard senses.
2. **LDA+Fixed** T: LDA with fixed T for each of nouns and verbs.
3. **LDA+Fixed** T**+Position**: LDA with fixed T and extra positional word features.
4. **HDP+Position**: HDP (which automatically learns T), with extra positional word features.
5. **HDP+Position+Dependency**: HDP with both positional word and dependency features.

We compare our models with two baselines from the SemEval-2010 task: (1) Baseline Random — randomly assign each test instance to one of four senses; (2) Baseline MFS — most frequent sense baseline, assigning all test instances to one sense; and also a benchmark system (**UoY**), in the form of the University of York system (Korkontzelos and Manandhar, 2010), which achieved the best overall WSD results in the original SemEval-2010 task.

3.2 SemEval-2010 Results

The results of our experiments over the SemEval-2010 dataset are summarised in Table 1.

System	WSD (80%/20%)		
	All	Verbs	Nouns
Baselines			
Baseline Random	0.57	0.66	0.51
Baseline MFS	0.59	0.67	0.53
LDA			
Variable T	0.64	0.69	0.60
Fixed T	0.63	0.68	0.59
Fixed T +Position	0.63	0.68	0.60
HDP			
+Position	0.68	0.72	0.65
+Position+Dependency	0.68	0.72	0.65
Benchmark			
UoY	0.62	0.67	0.59

Table 1: WSD F-score over the SemEval-2010 dataset

Looking first at the results for LDA, we see that the first LDA approach (variable T) is very competitive, outperforming the benchmark system. In this approach, however, we assume perfect knowledge of the number of gold senses of each target word, meaning that the method isn't truly unsupervised. When we fixed T for each of the nouns and verbs, we see a small drop in F-score, but encouragingly the method still performs above the benchmark. Adding positional word features improves the results very slightly for nouns.

When we relax the assumption on the number of word senses in moving to HDP, we observe a marked improvement in F-score over LDA. This is highly encouraging and somewhat surprising, as in hiding information about sense granularity from the model, we have actually *improved* our results. We return to discuss this effect below. For the final feature, we add dependency features to the HDP model (in addition to retaining the positional word features), but see no movement in the results.[6] While the dependency features didn't reduce F-score, their utility is questionable as the generation of the features from the Stanford parser is computationally expensive.

To better understand these results, we present the top-10 terms for each of the senses induced for the word *cheat* in Table 2. These senses are learnt using HDP with both positional word features (e.g. *husband_#-1*, indicating the lemma *husband* to the immediate left of the target word) and dependency features (e.g. *cheat#prep_on#wife*). The first observation to make is that senses 7, 8 and 9 are "junk" senses, in that the top-10 terms do not convey a coherent sense. These topics are an artifact of HDP: they are learnt at a much later stage of the iterative process of Gibbs sampling and are often smaller than other topics (i.e. have more zero-probability terms). We notice that they are assigned as topics to instances very rarely (although they are certainly used to assign topics to non-target *words* in the instances), and as such, they do not present a real issue when assigning the sense to an instance, as they are likely to be overshadowed by the dominant senses.[7] This conclusion is born out when we experimented with manually filtering out these topics when assigning instance to senses: there was no perceptible change in the results, reinforcing our suggestion that these topics do not impact on target word sense assignment.

Comparing the results for HDP back to those for LDA, HDP tends to learn almost double the number of senses per target word as are in the gold-standard (and hence are used for the "Variable T" version of LDA). Far from hurting our WSD F-score, however, the extra topics are dominated by junk topics, and boost WSD F-score for the "genuine" topics. Based on this insight, we ran LDA once again with variable T (and positional and dependency features), but this time setting T to the value learned by HDP, to give LDA the facility to use junk topics. This resulted in an F-score of 0.66 across all word classes (verbs = 0.71, nouns = 0.62), demonstrating that, surprisingly, even for the same T setting, HDP achieves superior results to LDA. I.e., not only does HDP learn T automatically, but the topic model learned for a given T is superior to that for LDA.

Looking at the other senses discovered for *cheat*, we notice that the model has induced a myriad of senses: the relationship sense of cheat (senses 1, 3 and 4, e.g. *husband cheats*); the exam usage of cheat (sense 2); the competition/game usage of cheat (sense 5); and cheating in the political domain (sense 6). Although the senses are possibly "split" a little more than desirable (e.g. senses 1, 3 and 4 arguably describe the same sense), the overall quality of the produced senses

[6] An identical result was observed for LDA.

[7] In the WSD evaluation, the alignment of induced senses to the gold senses is learnt automatically based on the mapping instances. E.g. if all instances that are assigned sense *a* have gold sense *x*, then sense *a* is mapped to gold sense *x*. Therefore, if the proportion of junk senses in the mapping instances is low, their influence on WSD results will be negligible.

Sense Num	Top-10 Terms
1	cheat think want ... love feel tell guy cheat#nsubj#include find
2	cheat student cheating test game school cheat#aux#to teacher exam study
3	husband wife cheat wife_#1 tiger husband_#-1 cheat#prep_on#wife ... woman cheat#nsubj#husband
4	cheat woman relationship cheating partner reason cheat#nsubj#man woman_#-1 cheat#aux#to spouse
5	cheat game play player cheating poker cheat#aux#to card cheated money
6	cheat exchange china chinese foreign cheat_#-2 cheat_#2 china_#-1 cheat#aux#to team
7	tina bette kirk walk accuse mon pok symkyn nick star
8	fat jones ashley pen body taste weight expectation parent able
9	euro goal luck fair france irish single 2000 cheat#prep_at#point complain

Table 2: The top-10 terms for each of the senses induced for the verb *cheat* by the HDP model (with positional word and dependency features)

is encouraging. Also, we observe a spin-off benefit of topic modelling approaches to WSI: the high-ranking words in each topic can be used to gist the sense, and anecdotally confirm the impact of the different feature types (i.e. the positional word and dependency features).

3.3 Comparison with other Topic Modelling Approaches to WSI

The idea of applying topic modelling to WSI is not entirely new. Brody and Lapata (2009) proposed an LDA-based model which assigns different weights to different feature sets (e.g. unigram tokens vs. dependency relations), using a "layered" feature representation. They carry out extensive parameter optimisation of both the (fixed) number of senses, number of layers, and size of the context window.

Separately, Yao and Durme (2011) proposed the use of non-parametric topic models in WSI. The authors preprocess the instances slightly differently, opting to remove the target word from each instance and stem the tokens. They also tuned the hyperparameters of the topic model to optimise the WSI effectiveness over the evaluation set, and didn't use positional or dependency features.

Both of these papers were evaluated over only the SemEval-2007 WSI dataset (Agirre and Soroa, 2007), so we similarly apply our HDP method to this dataset for direct comparability. In the remainder of this section, we refer to Brody and Lapata (2009) as BL, and Yao and Durme (2011) as YVD.

The SemEval-2007 dataset consists of roughly 27K instances, for 65 target verbs and 35 target nouns. BL report on results only over the noun instances, so we similarly restrict our attention to the nouns in this paper. Training data was not provided as part of the original dataset, so we follow the approach of BL and YVD in constructing our own training dataset for each target word from instances extracted from the British National Corpus (BNC: Burnard (2000)).[8] Both BL and YVD separately report slightly higher in-domain results from training on WSJ data (the SemEval-2007 data was taken from the WSJ). For the purposes of model comparison under identical training settings, however, it is appropriate to report on results for only the BNC.

System	F-Score
BL	0.855
YVD	0.857
SemEval Best (I2R)	0.868
Our method (default parameters)	0.842
Our method (tuned parameters)	0.869

Table 3: F-score for the SemEval-2007 WSI task, for our HDP method with default and tuned parameter settings, as compared to competitor topic modelling and other approaches to WSI

We experiment with both our original method (with both positional word and dependency features, and default parameter settings for HDP) without any parameter tuning, and the same method with the tuned parameter settings of YVD, for direct comparability. We present the results in Table 3, including the results for the best-performing system in the original SemEval-2007 task (I2R: Niu et al. (2007)).

The results are enlightening: with default parameter settings, our methodology is slightly below the results of the other three models. Bear

[8] In creating the training dataset, each instance is made up of the sentence the target word occurs in, as we as one sentence to either side of that sentence, i.e. 3 sentences in total per instance.

in mind, however, that the two topic modelling-based approaches were tuned extensively to the dataset. When we use the tuned hyperparameter settings of YVD, our results rise around 2.5% to surpass both topic modelling approaches, and marginally outperform the I2R system from the original task. Recall that both BL and YVD report higher results again using in-domain training data, so we would expect to see further gains again over the I2R system in following this path.

Overall, these results agree with our findings over the SemEval-2010 dataset (Section 3.2), underlining the viability of topic modelling to automated word sense induction.

3.4 Discussion

As part of our preprocessing, we remove all stopwords (other than for the positional word and dependency features), as described in Section 2.1. We separately experimented with not removing stopwords, based on the intuition that prepositions such as *to* and *on* can be informative in determining word sense based on local context. The results were markedly worse, however. We also tried appending part of speech information to each word lemma, but the resulting data sparseness meant that results dropped marginally.

When determining the sense for an instance, we aggregate the sense assignments for each word in the instance (not just the target word). An alternate strategy is to use only the target word topic assignment, but again, the results for this strategy were inferior to the aggregate method.

In the SemEval-2007 experiments (Section 3.3), we found that YVD's hyperparameter settings yielded better results than the default settings. We experimented with parameter tuning over the SemEval-2010 dataset (including YVD's optimal setting on the 2007 dataset), but found that the default setting achieved the best overall results: although the WSD F-score improved a little for nouns, it worsened for verbs. This observation is not unexpected: as the hyperparameters were optimised for nouns in their experiments, the settings might not be appropriate for verbs. This also suggests that their results may be due in part to overfitting the SemEval-2007 data.

4 Identifying Novel Senses

Having established the effectiveness of our approach at WSI, we next turn to an application of WSI, in identifying words which have taken on novel senses over time, based on analysis of diachronic data. Our topic modelling approach is particularly attractive for this task as, not only does it jointly perform type-level WSI, and token-level WSD based on the induced senses (in assigning topics to each instance), but it is possible to gist the induced senses via the contents of the topic (typically using the topic words with highest marginal probability).

The meanings of words can change over time; in particular, words can take on new senses. Contemporary examples of new word-senses include the meanings of *swag* and *tweet* as used below:

1. *We all know Frankie is adorable, but does he have swag?* [*swag* = 'style']
2. *The alleged victim gave a description of the man on Twitter and tweeted that she thought she could identify him.* [*tweet* = 'send a message on Twitter']

These senses of *swag* and *tweet* are not included in many dictionaries or computational lexicons — e.g., neither of these senses is listed in Wordnet 3.0 (Fellbaum, 1998) — yet appear to be in regular usage, particularly in text related to pop culture and online media.

The manual identification of such new word-senses is a challenge in lexicography over and above identifying new words themselves, and is essential to keeping dictionaries up-to-date. Moreover, lexicons that better reflect contemporary usage could benefit NLP applications that use sense inventories.

The challenge of identifying changes in word sense has only recently been considered in computational linguistics. For example, Sagi et al. (2009), Cook and Stevenson (2010), and Gulordava and Baroni (2011) propose type-based models of semantic change. Such models do not account for polysemy, and appear best-suited to identifying changes in predominant sense. Bamman and Crane (2011) use a parallel Latin–English corpus to induce word senses and build a WSD system, which they then apply to study diachronic variation in word senses. Crucially, in this token-based approach there is a clear connection between word senses and tokens, making it possible to identify usages of a specific sense.

Based on the findings in Section 3.2, here we apply the HDP method for WSI to the task of

identifying new word-senses. In contrast to Bamman and Crane (2011) our token-based approach does not require parallel text to induce senses.

4.1 Method

Given two corpora — a reference corpus which we take to represent standard usage, and a second corpus of newer texts — we identify senses that are novel to the second corpus compared to the reference corpus. For a given word w, we pool all usages of w in the reference corpus and second corpus, and run the HDP WSI method on this super-corpus to induce the senses of w. We then tag all usages of w in both corpora with their single most-likely automatically-induced sense.

Intuitively, if a word w is used in some sense s in the second corpus, and w is never used in that sense in the reference corpus, then w has acquired a new sense, namely s. We capture this intuition into a novelty score ("Nov") that indicates whether a given word w has a new sense in the second corpus, s, compared to the reference corpus, r, as below:

$$\text{Nov}(w) = \max\left(\left\{\frac{p_s(t_i) - p_r(t_i)}{p_r(t_i)} : t_i \in T\right\}\right) \quad (1)$$

where $p_s(t_i)$ and $p_r(t_i)$ are the probability of sense t_i in the second corpus and reference corpus, respectively, calculated using smoothed maximum likelihood estimates, and T is the set of senses induced for w. Novelty is high if there is some sense t that has much higher relative frequency in s than r and that is also relatively infrequent in r.

4.2 Data

Because we are interested in the identification of novel word-senses for applications such as lexicon maintenance, we focus on relatively newly-coined word-senses. In particular, we take the written portion of the BNC — consisting primarily of British English text from the late 20th century — as our reference corpus, and a similarly-sized random sample of documents from the ukWaC (Ferraresi et al., 2008) — a Web corpus built from the `.uk` domain in 2007 which includes a wide range of text types — as our second corpus. Text genres are represented to different extents in these corpora with, for example, text types related to the Internet being much more common in the ukWaC. Such differences are a noted challenge for approaches to identifying lexical semantic differences between corpora (Peirsman et al., 2010), but are difficult to avoid given the corpora that are available. We use TreeTagger (Schmid, 1994) to tokenise and lemmatise both corpora.

Evaluating approaches to identifying semantic change is a challenge, particularly due to the lack of appropriate evaluation resources; indeed, most previous approaches have used very small datasets (Sagi et al., 2009; Cook and Stevenson, 2010; Bamman and Crane, 2011). Because this is a preliminary attempt at applying WSI techniques to identifying new word-senses, our evaluation will also be based on a rather small dataset.

We require a set of words that are known to have acquired a new sense between the late 20th and early 21st centuries. The Concise Oxford English Dictionary aims to document contemporary usage, and has been published in numerous editions including Thompson (1995, COD95) and Soanes and Stevenson (2008, COD08). Although some of the entries have been substantially revised between editions, many have not, enabling us to easily identify new senses amongst the entries in COD08 relative to COD95. A manual linear search through the entries in these dictionaries would be very time consuming, but by exploiting the observation that new words often correspond to concepts that are culturally salient (Ayto, 2006), we can quickly identify some candidates for words that have taken on a new sense.

Between the time periods of our two corpora, computers and the Internet have become much more mainstream in society. We therefore extracted all entries from COD08 containing the word *computing* (which is often used as a topic label in this dictionary) that have a token frequency of at least 1000 in the BNC. We then read the entries for these 87 lexical items in COD95 and COD08 and identified those which have a clear computing sense in COD08 that was not present in COD95. In total we found 22 such items. This process, along with all the annotation in this section, is carried out by a native English-speaking author of this paper.

To ensure that the words identified from the dictionaries do in fact have a new sense in the ukWaC sample compared to the BNC, we examine the usage of these words in the corpora. We extract a random sample of 100 usages of each

lemma from the BNC and ukWaC sample and annotate these usages as to whether they correspond to the novel sense or not. This binary distinction is easier than fine-grained sense annotation, and since we do not use these annotations for formal evaluation — only for selecting items for our dataset — we do not carry out an inter-annotator agreement study here. We eliminate any lemma for which we find evidence of the novel sense in the BNC, or for which we do not find evidence of the novel sense in the ukWaC sample.[9] We further check word sketches (Kilgarriff and Tugwell, 2002)[10] for each of these lemmas in the BNC and ukWaC for collocates that likely correspond to the novel sense; we exclude any lemma for which we find evidence of the novel sense in the BNC, or fail to find evidence of the novel sense in the ukWaC sample. At the end of this process we have identified the following 5 lemmas that have the indicated novel senses in the ukWaC compared to the BNC: *domain* (n) "Internet domain"; *export* (v) "export data"; *mirror* (n) "mirror website"; *poster* (n) "one who posts online"; and *worm* (n) "malicious program". For each of the 5 lemmas with novel senses, a second annotator — also a native English-speaking author of this paper — annotated the sample of 100 usages from the ukWaC. The observed agreement and unweighted Kappa between the two annotators is 97.2% and 0.92, respectively, indicating that this is indeed a relatively easy annotation task. The annotators discussed the small number of disagreements to reach consensus.

For our dataset we also require items that have *not* acquired a novel sense in the ukWaC sample. For each of the above 5 lemmas we identified a distractor lemma of the same part-of-speech that has a similar frequency in the BNC, and that has not undergone sense change between COD95 and COD08. The 5 distractors are: *cinema* (n); *guess* (v); *symptom* (n); *founder* (n); and *racism* (n).

4.3 Results

We compute novelty ("Nov", Equation 1) for all 10 items in our dataset, based on the output of the topic modelling. The results are shown in column "Novelty" in Table 4. The lemmas with a novel sense have higher novelty scores than the distractors according to a one-sided Wilcoxon rank sum test ($p < .05$).

Lemma	Novelty	Freq. ratio	Novel sense freq.
domain (n)	116.2	2.60	41
worm (n)	68.4	1.04	30
mirror (n)	38.4	0.53	10
guess (v)	16.5	0.93	–
export (v)	13.8	0.88	28
founder (n)	11.0	1.20	–
cinema (n)	9.7	1.30	–
poster (n)	7.9	1.83	4
racism (n)	2.4	0.98	–
symptom (n)	2.1	1.16	–

Table 4: Novelty score ("Nov"), ratio of frequency in the ukWaC sample and BNC, and frequency of the novel sense in the manually-annotated 100 instances from the ukWaC sample (where applicable), for all lemmas in our dataset. Lemmas shown in boldface have a novel sense in the ukWaC sample compared to the BNC.

When a lemma takes on a new sense, it might also increase in frequency. We therefore also consider a baseline in which we rank the lemmas by the ratio of their frequency in the second and reference corpora. These results are shown in column "Freq. ratio" in Table 4. The difference between the frequency ratios for the lemmas with a novel sense, and the distractors, is not significant ($p > .05$).

Examining the frequency of the novel senses — shown in column "Novel sense freq." in Table 4 — we see that the lowest-ranked lemma with a novel sense, *poster*, is also the lemma with the least-frequent novel sense. This result is unsurprising as our novelty score will be higher for higher-frequency novel senses. The identification of infrequent novel senses remains a challenge.

The top-ranked topic words for the sense corresponding to the maximum in Equation 1 for the highest-ranked distractor, *guess*, are the following: @card@, *post*, ..., *n't*, *comment*, *think*, *subject*, *forum*, *view*, *guess*. This sense seems to correspond to usages of *guess* in the context of online forums, which are better represented in the ukWaC sample than the BNC. Because of the challenges posed by such differences between corpora (discussed in Section 4.2) we are unsurprised to see such an error, but this could be addressed in the future by building comparable cor-

[9] We use the IMS Open Corpus Workbench (`http://cwb.sourceforge.net/`) to extract the usages of our target lemmas from the corpora. This extraction process fails in some cases, and so we also eliminate such items from our dataset.

[10] `http://www.sketchengine.co.uk/`

Lemma	Topic Selection Methodology								
	Nov			Oracle (single topic)			Oracle (multiple topics)		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
domain (n)	1.00	0.29	0.45	1.00	0.56	0.72	0.97	0.88	0.92
export (v)	0.93	0.96	0.95	0.93	0.96	0.95	0.90	1.00	0.95
mirror (n)	0.67	1.00	0.80	0.67	1.00	0.80	0.67	1.00	0.80
poster (n)	0.00	0.00	0.00	0.44	1.00	0.62	0.44	1.00	0.62
worm (n)	0.93	0.90	0.92	0.93	0.90	0.92	0.86	1.00	0.92

Table 5: Results for identifying the gold-standard novel senses based on the three topic selection methodologies of: (1) Nov; (2) oracle selection of a single topic; and (3) oracle selection of multiple topics.

pora for use in this application.

Having demonstrated that our method for identifying novel senses can distinguish lemmas that have a novel sense in one corpus compared to another from those that do not, we now consider whether this method can also automatically identify the *usages* of the induced novel sense.

For each lemma with a gold-standard novel sense, we define the automatically-induced novel sense to be the single sense corresponding to the maximum in Equation 1. We then compute the precision, recall, and F-score of this novel sense with respect to the gold-standard novel sense, based on the 100 annotated tokens for each of the 5 lemmas with a novel sense. The results are shown in the first three numeric columns of Table 5.

In the case of *export* and *worm* the results are remarkably good, with precision and recall both over 0.90. For *domain*, the low recall is a result of the majority of usages of the gold-standard novel sense ("Internet domain") being split across two induced senses — the top-two highest ranked induced senses according to Equation 1. The poor performance for *poster* is unsurprising due to the very low frequency of this lemma's gold-standard novel sense.

These results are based on our novelty ranking method ("Nov"), and the assumption that the novel sense will be represented in a single topic. To evaluate the theoretical upper-bound for a topic-ranking method which uses our HDP-based WSI method and selects a single topic to capture the novel sense, we next evaluate an optimal topic selection approach. In the middle three numeric columns of Table 5, we present results for an experimental setup in which the single best induced sense — in terms of F-score — is selected as the novel sense by an oracle. We see big improvements in F-score for *domain* and *poster*. This encouraging result suggests refining the sense selection heuristic could theoretically improve our method for identifying novel senses, and that the topic modelling approach proposed in this paper has considerable promise for automatic novel sense detection. Of particular note is the result for *poster*: although the gold-standard novel sense of *poster* is rare, all of its usages are grouped into a single topic.

Finally, we consider whether an oracle which can select the best subset of induced senses — in terms of F-score — as the novel sense could offer further improvements. In this case — results shown in the final three columns of Table 5 — we again see an increase in F-score to 0.92 for *domain*. For this lemma the gold-standard novel sense usages were split across multiple induced topics, and so we are unsurprised to find that a method which is able to select multiple topics as the novel sense performs well. Based on these findings, in future work we plan to consider alternative formulations of novelty.

5 Conclusion

We propose the application of topic modelling to the task of word sense induction (WSI), starting with a simple LDA-based methodology with a fixed number of senses, and culminating in a nonparametric method based on a Hierarchical Dirichlet Process (HDP), which automatically learns the number of senses for a given target word. Our HDP-based method outperforms all methods over the SemEval-2010 WSI dataset, and is also superior to other topic modelling-based approaches to WSI based on the SemEval-2007 dataset. We applied the proposed WSI model to the task of identifying words which have taken on new senses, including identifying the token occurrences of the new word sense. Over a small dataset developed in this research, we achieved highly encouraging results.

References

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic.

John Ayto. 2006. *Movers and Shakers: A Chronology of Words that Shaped our Age*. Oxford University Press, Oxford.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 2011 Joint International Conference on Digital Libraries (JCDL 2011)*, pages 1–10, Ottawa, Canada.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

S. Brody and M. Lapata. 2009. Bayesian word sense induction. pages 103–111, Athens, Greece.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Paul Cook 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 (LREC 2010)*, pages 28–34, Valletta, Malta.

Marie-Catherine De Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. Genoa, Italy.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop: Can we beat Google*, pages 47–54, Marrakech, Morocco.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. pages 233–237.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, Scotland.

Adam Kilgarriff and David Tugwell. 2002. Sketching words. In Marie-Hélène Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137. Euralex, Grenoble, France.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pages 3–10, Whistler, Canada.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Uoy: Graphs of unambiguous vertices for word sense induction and disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 355–358, Uppsala, Sweden.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden.

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 116–126, Cambridge, USA.

Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. 2007. I2R: Three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 177–182, Prague, Czech Republic.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Comput. Linguist.*, 33:161–199.

Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Hinrich Schutze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Catherine Soanes and Angus Stevenson, editors. 2008. *The Concise Oxford English Dictionary*. Oxford University Press, eleventh (revised) edition. Oxford Reference Online.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Della Thompson, editor. 1995. *The Concise Oxford Dictionary of Current English*. Oxford University Press, Oxford, ninth edition.

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