Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models

Jey Han Lau,[♠] **Paul Cook,**[♡] **Diana McCarthy,**[♢] **Spandana Gella,**[♡] and **Timothy Baldwin**[♡]
♠ Dept of Philosophy, King's College London
♡ Dept of Computing and Information Systems, The University of Melbourne
♢ University of Cambridge
jeyhan.lau@gmail.com, paulcook@unimelb.edu.au,
diana@dianamccarthy.co.uk, spandanagella@gmail.com, tb@ldwin.net

Abstract

Unsupervised word sense disambiguation (WSD) methods are an attractive approach to all-words WSD due to their non-reliance on expensive annotated data. Unsupervised estimates of sense frequency have been shown to be very useful for WSD due to the skewed nature of word sense distributions. This paper presents a fully unsupervised topic modelling-based approach to sense frequency estimation, which is highly portable to different corpora and sense inventories, in being applicable to any part of speech, and not requiring a hierarchical sense inventory, parsing or parallel text. We demonstrate the effectiveness of the method over the tasks of predominant sense learning and sense distribution acquisition, and also the novel tasks of detecting senses which aren't attested in the corpus, and identifying novel senses in the corpus which aren't captured in the sense inventory.

1 Introduction

The automatic determination of word sense information has been a long-term pursuit of the NLP community (Agirre and Edmonds, 2006; Navigli, 2009). Word sense distributions tend to be Zipfian, and as such, a simple but surprisingly high-accuracy back-off heuristic for word sense disambiguation (WSD) is to tag each instance of a given word with its predominant sense (McCarthy et al., 2007). Such an approach requires knowledge of predominant senses; however, word sense distributions — and predominant senses too — vary from corpus to corpus. Therefore, methods for automatically learning predominant senses and sense distributions for specific corpora are required (Koeling et al., 2005; Lapata and Brew, 2004).

In this paper, we propose a method which uses topic models to estimate word sense distributions. This method is in principle applicable to all parts of speech, and moreover does not require a parser, a hierarchical sense representation or parallel text. Topic models have been used for WSD in a number of studies (Boyd-Graber et al., 2007; Li et al., 2010; Lau et al., 2012; Preiss and Stevenson, 2013; Cai et al., 2007; Knopp et al., 2013), but our work extends significantly on this earlier work in focusing on the acquisition of prior word sense distributions (and predominant senses).

Because of domain differences and the skewed nature of word sense distributions, it is often the case that some senses in a sense inventory will not be attested in a given corpus. A system capable of automatically finding such senses could reduce ambiguity, particularly in domain adaptation settings, while retaining rare but nevertheless viable senses. We further propose a method for applying our sense distribution acquisition system to the task of finding unattested senses — i.e., senses that are in the sense inventory but not attested in a given corpus. In contrast to the previous work of McCarthy et al. (2004a) on this topic which uses the sense ranking score from McCarthy et al. (2004b) to remove low-frequency senses from WordNet, we focus on finding senses that are unattested in the corpus on the premise that, given accurate disambiguation, rare senses in a corpus contribute to correct interpretation.

Corpus instances of a word can also correspond to senses that are not present in a given sense inventory. This can be due to, for example, words taking on new meanings over time (e.g. the rela-

Lau, Jey Han, Paul Cook, Diana McCarthy, Spandana Gella and Timothy Baldwin (to appear) 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 (ACL 2014), Baltimore, USA.

tively recent senses of *tablet* and *swipe* related to touchscreen computers) or domain-specific terms not being included in a more general-purpose sense inventory. A system for automatically identifying such novel senses — i.e. senses that are attested in the corpus but not in the sense inventory — would be a very valuable lexicographical tool for keeping sense inventories up-to-date (Cook et al., 2013). We further propose an application of our proposed method to the identification of such novel senses. In contrast to McCarthy et al. (2004b), the use of topic models makes this possible, using topics as a proxy for sense (Brody and Lapata, 2009; Yao and Durme, 2011; Lau et al., 2012). Earlier work on identifying novel senses focused on individual tokens (Erk, 2006), whereas our approach goes further in identifying groups of tokens exhibiting the same novel sense.

2 Background and Related Work

There has been a considerable amount of research on representing word senses and disambiguating usages of words in context (WSD) as, in order to produce computational systems that understand and produce natural language, it is essential to have a means of representing and disambiguating word sense. WSD algorithms require word sense information to disambiguate token instances of a given ambiguous word, e.g. in the form of sense definitions (Lesk, 1986), semantic relationships (Navigli and Velardi, 2005) or annotated data (Zhong and Ng, 2010). One extremely useful piece of information is the word sense prior or expected word sense frequency distribution. This is important because word sense distributions are typically skewed (Kilgarriff, 2004), and systems do far better when they take bias into account (Agirre and Martinez, 2004).

Typically, word frequency distributions are estimated with respect to a sense-tagged corpus such as SemCor (Miller et al., 1993), a 220,000 word corpus tagged with WordNet (Fellbaum, 1998) senses. Due to the expense of hand tagging, and sense distributions being sensitive to domain and genre, there has been some work on trying to estimate sense frequency information automatically (McCarthy et al., 2004b; Chan and Ng, 2005; Mohammad and Hirst, 2006; Chan and Ng, 2006). Much of this work has been focused on ranking word senses to find the predominant sense in a given corpus (McCarthy et al., 2004b; Mohammad and Hirst, 2006), which is a very powerful heuristic approach to WSD. Most WSD systems rely upon this heuristic for back-off in the absence of strong contextual evidence (McCarthy et al., 2007). McCarthy et al. (2004b) proposed a method which relies on distributionally similar words (nearest neighbours) associated with the target word in an automatically acquired thesaurus (Lin, 1998). The distributional similarity scores of the nearest neighbours are associated with the respective target word senses using a WordNet similarity measure, such as those proposed by Jiang and Conrath (1997) and Banerjee and Pedersen (2002). The word senses are ranked based on these similarity scores, and the most frequent sense is selected for the corpus that the distributional similarity thesaurus was trained over.

As well as sense ranking for predominant sense acquisition, automatic estimates of sense frequency distribution can be very useful for WSD for training data sampling purposes (Agirre and Martinez, 2004), entropy estimation (Jin et al., 2009), and prior probability estimates, all of which can be integrated within a WSD system (Chan and Ng, 2005; Chan and Ng, 2006; Lapata and Brew, 2004). Various approaches have been adopted, such as normalizing sense ranking scores to obtain a probability distribution (Jin et al., 2009), using subcategorisation information as an indication of verb sense (Lapata and Brew, 2004) or alternatively using parallel text (Chan and Ng, 2005; Chan and Ng, 2006; Agirre and Martinez, 2004).

The work of Boyd-Graber and Blei (2007) is highly related in that it extends the method of McCarthy et al. (2004b) to provide a generative model which assumes the words in a given document are generated according to the topic distribution appropriate for that document. They then predict the most likely sense for each word in the document based on the topic distribution and the words in context ("corroborators"), each of which, in turn, depends on the document's topic distribution. Using this approach, they get comparable results to McCarthy et al. when context is ignored (i.e. using a model with one topic), and at most a 1% improvement on SemCor when they use more topics in order to take context into account. Since the results do not improve on McCarthy et al. as regards sense distribution acquisition irrespective of context, we will compare our model with that proposed by McCarthy et al.

Recent work on finding novel senses has tended to focus on comparing diachronic corpora (Sagi et al., 2009; Cook and Stevenson, 2010; Gulordava and Baroni, 2011) and has also considered topic models (Lau et al., 2012). In a similar vein, Peirsman et al. (2010) considered the identification of words having a sense particular to one language variety with respect to another (specifically Belgian and Netherlandic Dutch). In contrast to these studies, we propose a model for comparing a corpus with a sense inventory. Carpuat et al. (2013) exploit parallel corpora to identify words in domain-specific monolingual corpora with previously-unseen translations; the method we propose does not require parallel data.

3 Methodology

Our methodology is based on the WSI system described in Lau et al. (2012),[1] which has been shown (Lau et al., 2012; Lau et al., 2013a; Lau et al., 2013b) to achieve state-of-the-art results over the WSI tasks from SemEval-2007 (Agirre and Soroa, 2007), SemEval-2010 (Manandhar et al., 2010) and SemEval-2013 (Navigli and Vannella, 2013; Jurgens and Klapaftis, 2013). The system is built around a Hierarchical Dirichlet Process (HDP: Teh et al. (2006)), a non-parametric variant of a Latent Dirichlet Allocation topic model (Blei et al., 2003) where the model automatically optimises the number of topics in a fully-unsupervised fashion over the training data.

To learn the senses of a target lemma, we train a single topic model per target lemma. The system reads in a collection of usages of that lemma, and automatically induces topics (= senses) in the form of a multinomial distribution over words, and per-usage topic assignments (= probabilistic sense assignments) in the form of a multinomial distribution over topics. Following Lau et al. (2012), we assign one topic to each usage by selecting the topic that has the highest cumulative probability density, based on the topic allocations of all words in the context window for that usage.[2] Note that in their original work, Lau et al. (2012) experimented with the use of features extracted from a dependency parser. Due to the computational overhead associated with these features, and the fact that the empirical impact of the features was found to be marginal, we make no use of parser-based features in this paper.[3]

The induced topics take the form of word multinomials, and are often represented by the top-N words in descending order of conditional probability. We interpret each topic as a sense of the target lemma.[4] To illustrate this, we give the example of topics induced by the HDP model for *network* in Table 1.

We refer to this method as HDP-WSI henceforth.[5]

In predominant sense acquisition, the task is to learn, for each target lemma, the most frequently occurring word sense in a particular domain or corpus, relative to a predefined sense inventory. The WSI system provides us with a topic allocation per usage of a given word, from which we can derive a distribution of topics over usages and a **predominant topic**. In order to map this onto the **predominant sense**, we need to have some way of aligning a topic with a sense. We design our topic–sense alignment methodology with portability in mind — it should be applicable to any sense inventory. As such, our alignment methodology assumes only that we have access to a conventional sense gloss or definition for each sense, and does not rely on ontological/structural knowledge (e.g. the WordNet hierarchy).

To compute the similarity between a sense and a topic, we first convert the words in the gloss/definition into a multinomial distribution over words, based on simple maximum likelihood estimation.[6] We then calculate the Jensen–Shannon divergence between the multinomial distribution (over words) of the gloss and that of the topic, and convert the divergence value into a similarity score by subtracting it from 1. Formally, the

[1] Based on the implementation available at: `https://github.com/jhlau/hdp-wsi`

[2] This includes all words in the usage sentence except stopwords, which were filtered in the preprocessing step.

[3] For hyper-parameters α and γ, we used 0.1 for both. We did not tune the parameters, and opted to use the default parameters introduced in Teh et al. (2006).

[4] To avoid confusion, we will refer to the HDP-induced topics as *topics*, and reserve the term *sense* to denote senses in a sense inventory.

[5] The code used to learn predominant sense and run all experiments described in this paper is available at: `https://github.com/jhlau/predom_sense`.

[6] Words are tokenised using OpenNLP and lemmatised with Morpha (Minnen et al., 2001). We additionally remove the target lemma, stopwords and words that are less than 3 characters in length.

Topic Num	Top-10 Terms
1	network support @card@ information research service group development community member
2	service @card@ road company transport rail area government network public
3	network social model system family structure analysis form relationship neural
4	network @card@ computer system service user access internet datum server
5	system network management software support corp company service application product
6	@card@ radio news television show bbc programme call think film
7	police drug criminal terrorist intelligence network vodafone iraq attack cell
8	network atm manager performance craigavon group conference working modelling assistant
9	root panos comenius etd unipalm lse brazil telephone xxx discuss

Table 1: An example to illustrate the topics induced for *network* by the HDP model. The top-10 highest probability terms are displayed to represent each topic (*@card@* denotes a tokenised cardinal number).

similarity sense s_i and topic t_j is:

$$\begin{aligned} \text{sim}(s_i, t_j) &= 1 - \text{JS}(S\|T) \\ &= 1 - \frac{1}{2}\text{KL}(S\|M) - \frac{1}{2}\text{KL}(T\|M) \end{aligned} \quad (1)$$

where: S and T are the multinomial distributions over words for sense s_i and topic t_j, respectively; $M = \frac{1}{2}(S+T)$; and $\text{JS}(X\|Y)$ and $\text{KL}(X\|Y)$ are the Jensen–Shannon and Kullback–Leibler divergence for distribution X and Y, respectively.

To learn the predominant sense, we compute the **prevalence score** of each sense and take the sense with the highest prevalence score as the predominant sense. The prevalence score for a sense is computed by summing the product of its similarity scores with each topic (i.e. $\text{sim}(s_i, t_j)$) and the prior probability of the topic in question (based on maximum likelihood estimation). Formally, the prevalence score of sense s_i is given as follows:

$$\begin{aligned} \text{prevalence}(s_i) &= \sum_j^T (\text{sim}(s_i, t_j) \times P(t_j)) \\ &= \sum_j^T \left(\text{sim}(s_i, t_j) \times \frac{f(t_j)}{\sum_k^T f(t_k)} \right) \end{aligned} \quad (2)$$

where $f(t_j)$ is the frequency of topic t_j (i.e. the number of usages assigned to topic t_j), and T is the number of topics.

The intuition behind the approach is that the predominant sense should be the sense that has relatively high similarity (in terms of lexical overlap) with high-probability topic(s).

4 WordNet Experiments

We first test the proposed method over the tasks of predominant sense learning and sense distribution induction, using the WordNet-tagged dataset of Koeling et al. (2005), which is made up of 3 collections of documents: a domain-neutral corpus (BNC), and two domain-specific corpora (SPORTS and FINANCE). For each domain, annotators were asked to sense-annotate a random selection of sentences for each of 40 target nouns, based on WordNet v1.7. The predominant sense and distribution across senses for each target lemma was obtained by aggregating over the sense annotations. The authors evaluated their method in terms of WSD accuracy over a given corpus, based on assigning all instances of a target word with the predominant sense learned from that corpus. For the remainder of the paper, we denote their system as MKWC.

To compare our system (HDP-WSI) with MKWC, we apply it to the three datasets of Koeling et al. (2005). For each dataset, we use HDP to induce topics for each target lemma, compute the similarity between the topics and the WordNet senses (Equation (1)), and rank the senses based on the prevalence scores (Equation (2)). In addition to the WSD accuracy based on the predominant sense inferred from a particular corpus, we additionally compute: (1) Acc_{UB}, the upper bound for the first sense-based WSD accuracy (using the gold standard predominant sense for disambiguation);[7] and (2) ERR, the error rate reduction between the accuracy for a given system (Acc) and the upper bound (Acc_{UB}), calculated as follows:

$$\text{ERR} = 1 - \frac{\text{Acc}_{\text{UB}} - \text{Acc}}{\text{Acc}_{\text{UB}}}$$

Looking at the results in Table 2, we see little difference in the results for the two methods, with MKWC performing better over two of the datasets (BNC and SPORTS) and HDP-WSI performing better over the third (FINANCE), but all

[7] The upper bound for a WSD approach which tags all token occurrences of a given word with the same sense, as a first step towards context-sensitive unsupervised WSD.

Dataset	FS_{CORPUS} Acc_{UB}	MKWC Acc	MKWC ERR	HDP-WSI Acc	HDP-WSI ERR
BNC	0.524	**0.407**	(0.777)	0.376	(0.718)
FINANCE	0.801	0.499	(0.623)	**0.555**	(0.693)
SPORTS	0.774	**0.437**	(0.565)	0.422	(0.545)

Table 2: WSD accuracy for MKWC and HDP-WSI on the WordNet-annotated datasets, as compared to the upper-bound based on actual first sense in the corpus (higher values indicate better performance; the **best** system in each row [other than the FS_{CORPUS} upper bound] is indicated in boldface).

differences are small. Based on the McNemar's Test with Yates correction for continuity, MKWC is significantly better over BNC and HDP-WSI is significantly better over FINANCE ($p < 0.0001$ in both cases), but the difference over SPORTS is not statistically significance ($p > 0.1$). Note that there is still much room for improvement with both systems, as we see in the gap between the upper bound (based on perfect determination of the first sense) and the respective system accuracies.

Given that both systems compute a continuous-valued prevalence score for each sense of a target lemma, a distribution of senses can be obtained by normalising the prevalence scores across all senses. The predominant sense learning task of McCarthy et al. (2007) evaluates the ability of a method to identify only the head of this distribution, but it is also important to evaluate the full sense distribution (Jin et al., 2009). To this end, we introduce a second evaluation metric: the Jensen–Shannon (JS) divergence between the inferred sense distribution and the gold-standard sense distribution, noting that smaller values are better in this case, and that it is now theoretically possible to obtain a JS divergence of 0 in the case of a perfect estimate of the sense distribution. Results are presented in Table 3.

HDP-WSI consistently achieves lower JS divergence, indicating that the distribution of senses that it finds is closer to the gold standard distribution. Testing for statistical significance over the paired JS divergence values for each lemma using the Wilcoxon signed-rank test, the result for FINANCE is significant ($p < 0.05$) but the results for the other two datasets are not ($p > 0.1$ in each case).

To summarise, the results for MKWC and HDP-WSI are fairly even for predominant sense learning (each outperforms the other at a level of statistical significance over one dataset), but HDP-WSI is better at inducing the overall sense distribution.

Dataset	MKWC	HDP-WSI
BNC	0.226	**0.214**
FINANCE	0.426	**0.375**
SPORTS	0.420	**0.363**

Table 3: Sense distribution evaluation of MKWC and HDP-WSI on the WordNet-annotated datasets, evaluated using JS divergence (lower values indicate better performance; the **best** system in each row is indicated in boldface).

Dataset	FS_{CORPUS} Acc_{UB}	FS_{DICT} Acc	FS_{DICT} ERR	HDP-WSI Acc	HDP-WSI ERR
UKWAC	0.574	0.387	(0.674)	**0.514**	(0.895)
TWITTER	0.468	0.297	(0.635)	**0.335**	(0.716)

Table 4: WSD accuracy for HDP-WSI on the Macmillan-annotated datasets, as compared to the upper-bound based on actual first sense in the corpus (higher values indicate better performance; the **best** system in each row [other than the FS_{CORPUS} upper bound] is indicated in boldface).

It is important to bear in mind that MKWC in these experiments makes use of full-text parsing in calculating the distributional similarity thesaurus, and the WordNet graph structure in calculating the similarity between associated words and different senses. Our method, on the other hand, uses no parsing, and only the synset definitions (and not the graph structure) of WordNet.[8] The non-reliance on parsing is significant in terms of portability to text sources which are less amenable to parsing (such as Twitter: (Baldwin et al., 2013)), and the non-reliance on the graph structure of WordNet is significant in terms of portability to conventional "flat" sense inventories. While comparable results on a different dataset have been achieved with a proximity thesaurus (McCarthy et al., 2007) compared to a dependency one,[9] it is not stated how wide a window is needed for the proximity thesaurus. This could be a significant issue with Twitter data, where context tends to be limited. In the

[8] McCarthy et al. (2004b) obtained good results with definition overlap, but their implementation uses the relation structure alongside the definitions (Banerjee and Pedersen, 2002). Iida et al. (2008) demonstrate that further extensions using distributional data are required when applying the method to resources without hierarchical relations.

[9] The thesauri used in the reimplementation of MKWC in this paper were obtained from `http://webdocs.cs.ualberta.ca/~lindek/downloads.htm`.

Dataset	FS_{CORPUS}	FS_{DICT}	HDP-WSI
UKWAC	0.210	0.393	**0.156**
TWITTER	0.259	0.472	**0.171**

Table 5: Sense distribution evaluation of HDP-WSI on the Macmillan-annotated datasets as compared to corpus- and dictionary-based first sense methods, evaluated using JS divergence (lower values indicate better performance; the **best** system in each row is indicated in boldface).

next section, we demonstrate the robustness of the method in experimenting with two new datasets, based on Twitter and a web corpus, and the Macmillan English Dictionary.

5 Macmillan Experiments

In our second set of experiments, we move to a new dataset (Gella et al., to appear) based on text from ukWaC (Ferraresi et al., 2008) and Twitter, and annotated using the Macmillan English Dictionary[10] (henceforth "Macmillan"). For the purposes of this research, the choice of Macmillan is significant in that it is a conventional dictionary with sense definitions and examples, but no linking between senses.[11] In terms of the original research which gave rise to the sense-tagged dataset, Macmillan was chosen over WordNet for reasons including: (1) the well-documented difficulties of sense tagging with fine-grained WordNet senses (Palmer et al., 2004; Navigli et al., 2007); (2) the regular update cycle of Macmillan (meaning it contains many recently-emerged senses); and (3) the finding in a preliminary sense-tagging task that it better captured Twitter usages than WordNet (and also OntoNotes: Hovy et al. (2006)).

The dataset is made up of 20 target nouns which were selected to span the high- to mid-frequency range in both Twitter and the ukWaC corpus, and have at least 3 Macmillan senses. The average sense ambiguity of the 20 target nouns in Macmillan is 5.6 (but 12.3 in WordNet). 100 usages of each target noun were sampled from each of Twitter (from a crawl over the time period Jan 3–Feb 28, 2013 using the Twitter Streaming API) and ukWaC, after language identification using `langid.py` (Lui and Baldwin, 2012) and POS tagging (based on the CMU ARK Twitter POS tagger v2.0 (Owoputi et al., 2012) for Twitter, and the POS tags provided with the corpus for ukWaC). Amazon Mechanical Turk (AMT) was then used to 5-way sense-tag each usage relative to Macmillan, including allowing the annotators the option to label a usage as "Other" in instances where the usage was not captured by any of the Macmillan senses. After quality control over the annotators/annotations (see Gella et al. (to appear) for details), and aggregation of the annotations into a single sense per usage (possibly "Other"), there were 2000 sense-tagged ukWaC sentences and Twitter messages over the 20 target nouns. We refer to these two datasets as UKWAC and TWITTER henceforth.

To apply our method to the two datasets, we use HDP-WSI to train a model for each target noun, based on the combined set of usages of that lemma in each of the two background corpora, namely the original Twitter crawl that gave rise to the TWITTER dataset, and all of ukWaC.

5.1 Learning Sense Distributions

As in Section 4, we evaluate in terms of WSD accuracy (Table 4) and JS divergence over the gold-standard sense distribution (Table 5). We also present the results for: (a) a supervised baseline ("FS_{CORPUS}"), based on the most frequent sense in the corpus; and (b) an unsupervised baseline ("FS_{DICT}"), based on the first-listed sense in Macmillan. In each case, the sense distribution is based on allocating all probability mass for a given word to the single sense identified by the respective method.

We first notice that, despite the coarser-grained senses of Macmillan as compared to WordNet, the upper bound WSD accuracy using Macmillan is comparable to that of the WordNet-based datasets over the balanced BNC, and quite a bit lower than that of the two domain corpora of Koeling et al. (2005). This suggests that both datasets are diverse in domain and content.

In terms of WSD accuracy, the results over UKWAC (ERR = 0.895) are substantially higher than those for BNC, while those over TWITTER (ERR = 0.716) are comparable. The accuracy is significantly higher than the dictionary-based first sense baseline (FS_{DICT}) over both datasets (McNemar's test; $p < 0.0001$), and the ERR is also considerably higher than for the two domain datasets in Section 4 (FINANCE and SPORTS). One cause of difficulty in sense-modelling TWITTER

[10] `http://www.macmillandictionary.com/`

[11] Strictly speaking, there is limited linking in the form of sets of synonyms in Macmillan, but we choose to not use this information in our research.

Dataset	P	R	F
UKWAC	0.73	0.85	0.74
TWITTER	0.56	0.88	0.65

Table 6: Evaluation of our method for identifying unattested senses, averaged over 10 runs of 10-fold cross validation

is large numbers of missing senses, with 12.3% of usages in TWITTER and 6.6% in UKWAC having no corresponding Macmillan sense.[12] This challenges the assumption built into the sense prevalence calculation that all topics will align to a pre-existing sense, a point we return to in Section 5.2.

The JS divergence results for both datasets are well below (= better than) the results for all three WordNet-based datasets, and also superior to both the supervised and unsupervised first-sense baselines. Part of the reason for this improvement is simply that the average polysemy in Macmillan (5.6 senses per target lemma) is slightly less than in WordNet (6.7 senses per target lemma),[13] making the task slightly easier in the Macmillan case.

5.2 Identification of Unattested Senses

We observed in Section 5.1 that there are relatively frequent occurrences of usages (e.g. 12.3% for TWITTER) which aren't captured by Macmillan. Conversely, there are also senses in Macmillan which aren't attested in the annotated sample of usages. Specifically, of the 112 senses defined for the 20 target lemmas, 25 (= 22.3%) of the senses are not attested in the 2000 usages in either corpora. Given that our methodology computes a prevalence score for each sense, it can equally be applied to the detection of these unattested senses, and it is this task that we address in this section: the identification of senses that are defined in the sense inventory but not attested in a given corpus.

Intuitively, an unused sense should have low similarity with the HDP induced topics. As such, we introduce sense-to-topic affinity, a measure that estimates how likely a sense is not attested in the corpus:

$$\text{st-affinity}(s_i) = \frac{\sum_j^T \text{sim}(s_i, t_j)}{\sum_k^S \sum_l^T \text{sim}(s_k, t_l)} \quad (3)$$

where $\text{sim}(s_i, t_j)$ is carried over from Equation (1), and T and S represent the number of topics and senses, respectively.

We treat the task of identification of unused senses as a binary classification problem, where the goal is to find a sense-to-topic affinity threshold below which a sense will be considered to be unused. We pool together all the senses and run 10-fold cross validation to learn the threshold for identifying unused senses,[14] evaluated using sense-level precision (P), recall (R) and F-score (F) at detecting unattested senses. We repeat the experiment 10 times (partitioning the items randomly into folds) and collect the mean precision, recall and F-scores across the 10 runs. We found encouraging results for the task, as detailed in Table 6. For the threshold, the average value with standard deviation is 0.092 ± 0.044 over UKWAC and 0.125 ± 0.052 over TWITTER, indicating relative stability in the value of the threshold both internally within a dataset, and also across datasets.

5.3 Identification of Novel Senses

In both TWITTER and UKWAC, we observed frequent occurrences of usages of our target nouns which didn't map onto a pre-existing Macmillan sense. A natural question to ask is whether our method can be used to predict word senses that are missing from our sense inventory, and identify usages associated with each such missing sense. We will term these "novel senses", and define "novel sense identification" to be the task of identifying new senses that are not recorded in the inventory but are seen in the corpus.

An immediate complication in evaluating novel sense identification is that we are attempting to identify senses which explicitly aren't in our sense inventory. This contrasts with the identification of unattested senses, e.g., where we were attempting to identify which of the *known* senses wasn't observed in the corpus. Also, while we have annotations of "Other" usages in TWITTER and UKWAC, there is no real expectation that all such usages will correspond to the same sense: in practice, they are attributable to a myriad of effects such as incorporation in a non-compositional multiword expression, and errors in POS tagging (i.e. the usage not being nominal). As such, we can't use the "Other" annotations to evaluate novel sense iden-

[12] The relative occurrence of unlisted/unclear senses in the datasets of Koeling et al. (2005) is comparable to UKWAC.

[13] Note that the set of lemmas differs between the respective datasets, so this isn't an accurate reflection of the relative granularity of the two dictionaries.

[14] We used a fixed step and increment at steps of 0.001, up to the max value of st-affinity when optimising the threshold.

No. Lemmas with a Removed Sense	Relative Freq of Removed Sense	Threshold Mean±stdev	P	R	F
20	0.0–0.2	0.052±0.009	0.35	0.42	0.36
9	0.2–0.4	0.089±0.024	0.24	0.59	0.29
6	0.4–0.6	0.061±0.004	0.63	0.64	0.63

Table 7: Classification of usages with novel sense for all target lemmas.

No. Lemmas with a Removed Sense	Relative Freq of Removed Sense	Threshold Mean±stdev	P	R	F
9	0.2–0.4	0.093±0.023	0.50	0.66	0.52
6	0.4–0.6	0.099±0.018	0.73	0.90	0.80

Table 8: Classification of usages with novel sense for target lemmas with a removed sense.

tification. The evaluation of systems for this task is a known challenge, which we address similarly to Erk (2006) by artificially synthesising novel senses through removal of senses from the sense inventory. In this way, even if we remove multiple senses for a given word, we still have access to information about which usages correspond to which novel sense. An additional advantage of this procedure is that it allows us to control an important property of novel senses: their frequency of occurrence.

In the experiments that follow, we randomly select senses for removal from three frequency bands: low, medium and high frequency senses. Frequency is defined by relative occurrence in the annotated usages: low = 0.0–0.2; medium = 0.2–0.4; and high = 0.4–0.6. Note that we do not consider high-frequency senses with frequency higher than 0.6, as it is rare for a medium- to high-frequency word to take on a novel sense which is then the predominant sense in a given corpus. Note also that not all target lemmas will have a novel sense through synthesis, as they may have no senses that fall within the indicated bounds of relative occurrence (e.g. if $> 60\%$ of usages are a single sense). For example, only 6 of our 20 target nouns have senses which are candidates for high-frequency novel senses.

As before, we treat the novel sense identification task as a classification problem, although with a significantly different formulation: we are no longer attempting to identify pre-existing senses, as novel senses are by definition not included in the sense inventory. Instead, we are seeking to identify clusters of usages which are instances of a novel sense, e.g. for presentation to a lexicographer as part of a dictionary update process (Rundell and Kilgarriff, 2011; Cook et al., 2013). That is, for each usage, we want to classify whether it is an instance of a given novel sense.

A usage that corresponds to a novel sense should have a topic that does not align well with any of the pre-existing senses in the sense inventory. Based on this intuition, we introduce topic-to-sense affinity to estimate the similarity of a topic to the set of senses, as follows:

$$\text{ts-affinity}(t_j) = \frac{\sum_i^S \text{sim}(s_i, t_j)}{\sum_l^T \sum_k^S \text{sim}(s_k, t_l)} \quad (4)$$

where, once again, $\text{sim}(s_i, t_j)$ is defined as in Equation (1), and T and S represent the number of topics and senses, respectively.

Using topic-to-sense affinity as the sole feature, we pool together all instances and optimise the affinity feature to classify instances that have novel senses. Evaluation is done by computing the mean precision, recall and F-score across 10 separate runs; results are summarised in Table 7. Note that we evaluate only over UKWAC in this section, for ease of presentation.

The results show that instances with high-frequency novel senses are more easily identifiable than instances with medium/low-frequency novel senses. This is unsurprising given that high-frequency senses have a higher probability of generating related topics (sense-related words are observed more frequently in the corpus), and as such are more easily identifiable.

We are interested in understanding whether pooling all instances — instances from target lemmas that have a sense artificially removed and those that do not — impacted the results (recall that not all target lemmas have a removed sense). To that end, we chose to include only instances from lemmas with a removed sense, and repeated the experiment for the medium- and

No. of Lemmas with a Removed Sense	No. of Lemmas without a Removed Sense	Relative Freq of Removed Sense	Wilcoxon Rank Sum p-value
10	0	0.0–0.2	0.4543
9	11	0.2–0.4	0.0391
6	14	0.4–0.6	0.0247

Table 9: Wilcoxon Rank Sum p-value results for testing target lemmas with removed sense vs. target lemmas without removed sense using novelty.

high-frequency novel sense condition (for the low-frequency condition, all target lemmas have a novel sense). In other words, we are assuming knowledge of which words have novel sense, and the task is to identify specifically what the novel sense is, as represented by novel usages. Results are presented in Table 8.

From the results, we see that the F-scores improved notably. This reveals that an additional step is necessary to determine whether a target lemma has a potential novel sense before feeding its instances to learn which of them contains the usage of the novel sense.

In the last experiment, we propose a new measure to tackle this: the identification of target lemmas that have a novel sense. We introduce novelty, a measure of the likelihood of a target lemma w having a novel sense:

$$\text{novelty}(w) = \min_{t_j} \left(\max_{s_i} \frac{\text{sim}(s_i, t_j)}{f(t_j)} \right) \quad (5)$$

where $f(t_j)$ is the frequency of topic t_j in the corpus. The intuition behind novelty is that a target lemma with a novel sense should have a (somewhat-)frequent topic that has low association with any sense. That we use the frequency rather than the probability of the topic here is deliberate, as topics with a higher raw number of occurrences (whether as a low-probability topic for a high-frequency word, or a high-probability topic for a low-frequency word) are indicative of a novel word sense.

For each of our three datasets (with low-, medium- and high-frequency novel senses, respectively), we compute the novelty of the target lemmas and the p-value of a one-tailed Wilcoxon rank sum test to test if the two groups of lemmas (i.e. lemmas with a novel sense vs. lemmas without a novel sense) are statistically different.[15] Results are presented in Table 9. We see that the novelty measure can readily identify target lemmas with high- and medium-frequency novel senses ($p < 0.05$), but the results are less promising for the low-frequency novel senses.

[15]Note that the number of words with low-frequency novel senses here is restricted to 10 (cf. 20 in Table 7) to ensure we have both positive and negative lemmas in the dataset.

6 Discussion

Our methodologies for the two proposed tasks of identifying unused and novel senses are simple extensions to demonstrate the flexibility and robustness of our methodology. Future work could pursue a more sophisticated methodology, using non-linear combinations of $\text{sim}(s_i, t_j)$ for computing the affinity measures or multiple features in a supervised context. We contend, however, that these extensions are ultimately a preliminary demonstration to the flexibility and robustness of our methodology.

A natural next step for this research would be to couple sense distribution estimation and the detection of unattested senses with evidence from the context, using topics or other information about the local context (e.g. Agirre and Soroa (2009)) to carry out unsupervised WSD of individual token occurrences of a given word.

In summary, we have proposed a topic modelling-based method for estimating word sense distributions, based on Hierarchical Dirichlet Processes and the earlier work of Lau et al. (2012) on word sense induction, in probabilistically mapping the automatically-learned topics to senses in a sense inventory. We evaluated the ability of the method to learn predominant senses and induce word sense distributions, based on a broad range of datasets and two separate sense inventories. In doing so, we established that our method is comparable to the approach of McCarthy et al. (2007) at predominant sense learning, and superior at inducing word sense distributions. We further demonstrated the applicability of the method to the novel tasks of detecting word senses which are unattested in a corpus, and identifying novel senses which are found in a corpus but not captured in a word sense inventory.

Acknowledgements

We wish to thank the anonymous reviewers for their valuable comments. This research was supported in part by funding from the Australian Research Council.

References

Eneko Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, Netherlands.

Eneko Agirre and David Martinez. 2004. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of EMNLP 2004*, pages 25–32, Barcelona, Spain.

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12, Prague, Czech Republic.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 33–41, Athens, Greece.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 136–145, Mexico City, Mexico.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David Blei. 2007. Putop: Turning predominant senses into a topic model for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 277–281, Prague, Czech Republic.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, Prague, Czech Republic.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 103–111, Athens, Greece.

Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. 2007. NUS-ML: Improving word sense disambiguation using topic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 249–252, Prague, Czech Republic.

Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. SenseSpotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1435–1445, Sofia, Bulgaria.

Yee Seng Chan and Hwee Tou Ng. 2005. Word sense disambiguation with distribution estimation. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1010–1015, Edinburgh, UK.

Yee Seng Chan and Hwee Tou Ng. 2006. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 89–96, Sydney, Australia.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 28–34, Valletta, Malta.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word senses. In *Proceedings of eLex 2013*, pages 49–65, Tallinn, Estonia.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 128–135, New York City, USA.

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

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.

Spandana Gella, Paul Cook, and Timothy Baldwin. to appear. One sense per tweeter ... and other lexical semantic tales of Twitter. In *Proceedings of the 14th Conference of the EACL (EACL 2014)*, Gothenburg, Sweden.

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

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 57–60, New York City, USA.

Ryu Iida, Diana McCarthy, and Rob Koeling. 2008. Gloss-based semantic similarity metrics for predominant sense acquisition. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 561–568.

Jay Jiang and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages 19–33, Taipei, Taiwan.

Peng Jin, Diana McCarthy, Rob Koeling, and John Carroll. 2009. Estimating and exploiting the entropy of sense distributions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies 2009 (NAACL HLT 2009): Short Papers*, pages 233–236, Boulder, USA.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the 7th International Workshop*

on Semantic Evaluation (SemEval 2013), pages 290–299, Atlanta, USA.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? Technical Report ITRI-04-10, Information Technology Research Institute, University of Brighton.

Johannes Knopp, Johanna Völker, and Simone Paolo Ponzetto. 2013. Topic modeling for word sense induction. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 97–103, Darmstadt, Germany.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 419–426, Vancouver, Canada.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–75.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France.

Jey Han Lau, Paul Cook, and Timothy Baldwin. 2013a. unimelb: Topic modelling-based word sense induction. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307–311, Atlanta, USA.

Jey Han Lau, Paul Cook, and Timothy Baldwin. 2013b. unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 217–221, Atlanta, USA.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Ontario, Canada.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147, Uppsala, Sweden.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 768–774, Montreal, Canada.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

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.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004a. Automatic identification of infrequent word senses. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 1220–1226, Geneva, Switzerland.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004b. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 280–287, Barcelona, Spain.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 4(33):553–590.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Saif Mohammad and Graeme Hirst. 2006. Determining word sense dominance using a thesaurus. In *Proceedings of EACL-2006*, pages 121–128, Trento, Italy.

Roberto Navigli and Daniele Vannella. 2013. SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, USA.

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1088.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 30–35, Prague, Czech Republic.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-speech tagging for Twitter: Word clusters and other advances. Technical Report CMU-ML-12-107, Machine Learning Department, Carnegie Mellon University.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *Proceedings of the HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 49–56, Boston, USA.

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

Judita Preiss and Mark Stevenson. 2013. Unsupervised domain tuning to improve word sense disambiguation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 680–684, Atlanta, USA.

Michael Rundell and Adam Kilgarriff. 2011. Automating the creation of dictionaries: where will it all end? In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot, editors, *A Taste for Corpora. In honour of Sylviane Granger*, pages 257–282. John Benjamins, Amsterdam, Netherlands.

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.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

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, USA.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden.