Novel Word-sense Identification

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

Abstract

Automatic lexical acquisition has been an active area of research in computational linguistics for over two decades, but the automatic identification of new word-senses has received attention only very recently. Previous work on this topic has been limited by the availability of appropriate evaluation resources. In this paper we present the largest corpus-based dataset of diachronic sense differences to date, which we believe will encourage further work in this area. We then describe several extensions to a state-of-the-art topic modelling approach for identifying new word-senses. This adapted method shows superior performance on our dataset of two different corpus pairs to that of the original method for both: (a) types having taken on a novel sense over time; and (b) the token instances of such novel senses.

1 Novel word-senses

The meanings of words change over time with, in particular, established words taking on new senses. For example, the usages of *drop*, *wall*, and *blow up* in the following sentences correspond to relatively-recent senses of these words that appear to be quite common in text related to popular culture, but are not listed in many dictionaries; for example, they are all missing from WordNet 3.0 (Fellbaum, 1998).

1. *The reissue album drops March 27 and is an extension of Perry's huge 2010 Teenage Dream.* [*drops* = "comes out", "is released"]

2. *On Facebook, you can plainly see much of the data the site has on you, because it's posted to your wall.* [*wall* = "Facebook wall", "personal electronic noticeboard"]

3. *Why would I give him my number so he can blow up my phone the way he does my inbox.* [*blow up* = "overwhelm with messages"]

Computational lexicons are an essential component of systems for a variety of natural language processing (NLP) tasks. The success of such systems, therefore, depends on the quality of the lexicons they use, and (semi-)automatic techniques for identifying new word-senses could benefit applied NLP by helping to keep lexicons up-to-date. In revising dictionaries, lexicographers must identify new word-senses, in addition to new words themselves; methods which identify new word-senses could therefore also help to keep dictionaries current.

In this paper, because of the need for lexicon maintenance, we focus on relatively-new word-senses. Specifically, we consider the identification of word-senses that are not attested in a *reference corpus*, taken to represent standard usage, but that are attested in a *focus corpus* of newer texts.

Lau et al. (2012) introduced the task of novel sense identification. They presented a method for identifying novel word-senses — described here in Section 4 — and evaluated this method on a very small dataset consisting of just five lemmas having a novel sense in a single corpus pair. Cook et al. (2013) extended the method of Lau et al. to incorporate knowledge of the expected domains of new word-senses, but did not conduct a rigorous empirical evaluation. The remainder of this paper is structured

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1624–1635, Dublin, Ireland, August 23-29 2014.

as follows. After discussing related work in Section 2, we present a substantially-expanded evaluation dataset in Section 3, that is based on a second corpus pair and consists of many more lemmas with a novel sense. We describe the models used by Lau et al. and Cook et al., and our new extensions to them, in Section 4. In Section 5 we analyse the results of novel sense identification, and consider a new baseline for this task. We demonstrate that the extended methods give an improvement over the original method of Lau et al. We conclude by discussing some previously-unexplored variations on novel sense identification, and limitations of the approaches considered.

The primary contributions of this paper are: (1) development of a novel sense detection dataset much larger than has been used in research to date; (2) development and evaluation of a new baseline for novel sense detection, reformulations of the method of Lau et al., and a method that incorporates only the expected domain(s) of novel senses; (3) empirical evaluation of the method of Cook et al.; and (4) extension of the novel sense detection method of Cook et al. to automatically acquire information about the expected domain(s) of novel senses.

2 Related work

Identifying diachronic changes in word-sense is a challenge that has only been considered rather recently in computational linguistics. Sagi et al. (2009) and Cook and Stevenson (2010) propose methods to identify specific types of semantic change — widening and narrowing, and amelioration and pejoration, respectively — based on specific properties of these phenomena. Gulordava and Baroni (2011) identify diachronic sense change in an n-gram database, but using a model that is not restricted to any particular type of semantic change. Cook and Hirst (2011) consider the impact of sense frequency on methods for identifying novel senses. Crucially, all of the aforementioned approaches are type-based: they are able to identify words that have undergone a change in meaning, but not the token instances which give rise to these sense differences.

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 sense frequency. Rohrdantz et al. (2011) present a system for visualizing changes in word usage over time. Crucially, in these token-based approaches there is a clear connection between (induced) word-senses and tokens, making it possible to identify usages of a specific (new) sense.

Other work has focused on sense differences between dialects and domains. Peirsman et al. (2010) consider the identification of words that are typical of Belgian and Netherlandic Dutch, due to either marked frequency or sense. McCarthy et al. (2007) consider the identification of predominant word-senses in corpora, including differences between domains. However, this approach does not identify new senses as it relies on a pre-existing sense inventory. Carpuat et al. (2013) identify words in a domain-specific parallel corpus with novel translations.

The method proposed by Lau et al. (2012), and extended by Cook et al. (2013), identifies novel word-senses using a state-of-the-art word-sense induction (WSI) system. This token-based approach offers a natural account of polysemy and not only identifies word types that have a novel sense, but identifies the token instances of the hypothesized novel senses, without reliance on parallel text or a pre-existing sense inventory. We therefore adopt this method for evaluation on our new dataset, and propose further extensions to this method.

3 Datasets

Evaluating approaches to identifying semantic change is a challenge due to the lack of appropriate evaluation resources (i.e., corpora for the appropriate time periods, known to exhibit particular sense changes); indeed, most previous approaches have used very small datasets (e.g., Sagi et al., 2009; Cook and Stevenson, 2010; Bamman and Crane, 2011). In this study we consider two datasets of relatively newly-coined word-senses: (1) an extended version of the dataset based on the BNC (Burnard, 2000) and ukWaC (Ferraresi et al., 2008) used by Lau et al. (2012); and (2) a new dataset based on the SiBol/Port Corpus.[1] This

[1] http://www3.lingue.unibo.it/blog/clb/?page_id=8

is the largest dataset for evaluating approaches to identifying diachronic semantic change constructed from corpus evidence to be presented to date.

3.1 BNC–ukWaC

Lau et al. (2012) take the written portion of the BNC (approximately 87 million words of British English from the late 20th century) as the reference corpus, and a similarly-sized random sample of documents from the ukWaC (a Web corpus built from the `.uk` domain in 2007) as the focus corpus. They used TreeTagger (Schmid, 1994) to tokenise and lemmatise both corpora.

A set of words that has acquired a new sense between the late 20th and early 21st centuries — the time periods of the reference and focus corpora — is required. 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 (2005, COD08), enabling the identification of new senses amongst the entries in COD08 relative to COD95. Manually searching these dictionaries for new senses would be time intensive, but new words often correspond to concepts that are culturally salient (Ayto, 2006), and one can leverage this observation to speed up the process of finding some candidate words with novel senses.[2]

Between the time periods of the reference and focus corpora, computers and the Internet have become much more mainstream in society. Lau et al. therefore extracted all headwords in COD08 whose entries contain the word *computing*. They then carefully annotated these lemmas to identify those that indeed exhibit the novel sense indicated in the dictionary in the corpora. Here, we expand Lau et al.'s dataset by extracting all headwords including any of the following words *code*, *computer*, *internet*, *network*, *online*, *program*, *web*, and *website*. We then follow a similar annotation process to Lau et al.

An annotator read the entries for the selected lexical items in COD95 and COD08, and identified those which have a clear sense related to computers or the Internet in COD08 that is not present in COD95; such senses are referred to as *novel senses*. This process, along with all the annotation in this section (including Section 3.2), is carried out by native English-speaking authors of this paper and graduate students in computational linguistics.

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 word sketches (Kilgarriff et al., 2004)[3] 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.[4]

We further 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 usages from the BNC, or for which we do not find evidence of the novel sense in the ukWaC sample usages.[5]

This process resulted in the identification of two lemmas not in the dataset of Lau et al., with frequency greater than 1000 in the ukWaC sample, and having a novel sense in the ukWaC compared to the BNC (*feed* (n) and *visit* (v)). Combining these new lemmas with the dataset of Lau et al. gives an expanded dataset consisting of seven lemmas. For both of the two new lemmas, a second annotator annotated the sample of 100 usages from the ukWaC. The observed agreement and unweighted Kappa for this annotation task for all seven lemmas is 97.4% and 0.93, respectively, indicating that this is indeed a relatively easy annotation task. The annotators discussed the small number of disagreements to reach

[2]We access the dictionaries in the same way as Lau et al., namely we use COD08 online via `http://oxfordreference.com`, and the paper version of COD95.

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

[4]We examine word sketches for the full ukWaC because this version of the corpus is available through the Sketch Engine.

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

BNC–ukWaC

Lemma	Frequency	Novel sense definition
domain (n)	41	Internet domain
export (v)	28	export data
feed (n)	23	data feed
mirror (n)	10	mirror website
poster (n)	4	one who posts online
visit (v)	28	access a website
worm (n)	30	malicious program

SiBol/Port

Lemma	Frequency	Novel sense definition
cloud (n)	9	Internet-based computational resources
drag (v)	1	move on a computer screen using a mouse
follower (n)	34	Twitter follower
help (n)	1	displayed instructions, e.g., help menu
hit (n)	2	search hit
platform (n)	22	computing platform
poster (n)	5	one who posts online
reader (n)	3	e-reader
rip (v)	1	copy music
site (n)	39	website
text (n)	39	text message
visit (v)	7	access a website
wall (n)	2	Facebook wall

Table 1: Lemmas in the BNC–ukWaC and SiBol/Port datasets. For each lemma, the frequency of its novel sense in the annotated sample of usages from the focus corpus, and a definition of its novel sense, are shown.

consensus. The seven lemmas in this dataset are shown in Table 1, along with definitions of their novel senses, and the frequencies of their novel senses in the focus corpus.

Lau et al. compared the novelty of the lemmas with a novel sense to that of a same-size set of distractor lemmas not having a novel sense. Here we consider a much larger set of 50 distractors — 25 nouns and 25 verbs — randomly sampled from a similar frequency range as the items with a novel sense.

One shortcoming of this dataset (and indeed the subset of it used by Lau et al.) is that text types are represented to different extents in the BNC and ukWaC, with, for example, texts related to the Internet being much more common in the ukWaC. Such differences in corpus composition are a noted challenge for approaches to identifying lexical semantic differences between corpora (Peirsman et al., 2010). In the following subsection we therefore consider the creation of a new dataset from more-comparable corpora.

3.2 SiBol/Port

The SiBol/Port Corpus consists of texts from several British newspapers for the years 1993, 2005, and 2010; we use the 1993 and 2010 portions of this corpus — referred to as SP1993 and SP2010 — as our reference and focus corpora, respectively. SP1993 and SP2010 contain approximately 93M and 99M words, respectively. In contrast to BNC–ukWaC, our reference and focus corpora are now comparable, in that they both consist of texts from British newspapers but they differ with respect to the specific year.

The novel word-senses in the BNC–ukWaC dataset are all related to computers and the Internet, but there has been recent lexical semantic change unrelated to technology as well (e.g., *sick* can be used to mean "excellent"). In an effort to include such non-technical novel senses in this new dataset, we obtain a list of headwords for which a sense was added to the Macmillan English Dictionary for Advanced

Learners (MEDAL)[6] since its first edition (Rundell and Fox, 2002), courtesy of Macmillan Dictionaries. Beginning with these candidates from MEDAL, and the items extracted from COD from Section 3.1, we discard any lemma whose frequency is less than 1000 in SP1993 or SP2010.

As for the BNC–ukWaC dataset, an annotator examined word sketches for these lemmas. However, it is possible that the novel sense for a lemma is present in a corpus, but that we fail to find evidence for it in that lemma's word sketch. We therefore also obtain judgements from two annotators as to whether each novel sense is expected to be very infrequent (or unattested) in SP2010. To reduce subsequent annotation effort, we discard any lemma for which its novel sense is believed to be infrequent in SP2010 by both judges, and is not found in the word sketch from SP2010.

Annotators then annotate a random sample of 100 usages of each lemma in the reference and focus corpora as before, and again eliminate any lemma for which we find evidence of its novel sense in the reference corpus, or fail to find evidence of that sense in the focus corpus. We identify thirteen lemmas having a novel sense in SP2010 relative to SP1993. These lemmas are also shown in Table 1.

We obtain a second set of annotations for the usages of these lemmas in the sample from SP2010, with each lemma being annotated by a different annotator than before. The observed agreement and unweighted Kappa between the two sets of annotations is 96.2% and 0.81, respectively. In cases of disagreement, a final annotation is again reached through discussion.

We randomly select 164 lemmas (116 nouns and 48 verbs) from a similar frequency range as the lemmas having a novel sense, to serve as distractors.

Both the BNC–ukWaC and SiBol/Port datasets have been made available.[7]

4 The WSI-based approach to novel word-sense detection

In this section we describe the WSI-based method of Lau et al. (2012) for detecting novel senses, and an extension of this method from Cook et al. (2013). We then present new extensions of this method.

The Lau et al. (2012) WSI model is based on a Hierarchical Dirichlet Process (HDP, Teh et al., 2006), which is a non-parametric variant of a topic model that, like the commonly-used Latent Dirichlet Allocation (LDA, Blei et al., 2003), learns topics (in the form of multinomial probability distributions over words) and per-document topic assignments (in the form of multinomial probability distributions over topics) for a collection of documents; unlike LDA, however, it also optimises the number of topics in an unsupervised data-driven manner. In the context of WSI, by creating "documents" that consist of sentences containing a target word, we can view the topics learnt by topic models as the sense representation of the target word. Indeed, topic models have been previously applied to WSI (e.g., Brody and Lapata, 2009; Yao and Van Durme, 2011).

To generate the input for the topic model, the documents are tokenised (in this case, a "document" is a short context, typically 1–3 sentences, containing a target word) into a bag of words. All words are lemmatised, and stopwords and low frequency terms are removed. Positional word features — commonly used in WSI — for each of the three words to the left and right of the target word are also included.

To induce the senses of a target word w from a given set of usages of w, HDP is run on those usages (represented according to the features described above) to induce topics; these topics are then interpreted as representing the senses of w (one topic per sense). To determine the sense assigned to each instance, the system aggregates over the topic assignments for each word in the context of w, and selects the topic with the highest aggregated probability, i.e., $\text{argmax}_z P(t = z|d)$, where d is a document and t is a topic.

Recently, Lau et al. (2013a,b) found this method to give the overall best performance on two WSI shared tasks (Jurgens and Klapaftis, 2013; Navigli and Vannella, 2013), demonstrating that the method is competitive with the state-of-the-art in WSI, and appropriate as the basis for a method for identifying novel word-senses.

[6] http://www.macmillandictionary.com/

[7] http://www.csse.unimelb.edu.au/~tim/etc/novel-sense-dataset.tgz

4.1 Novel Sense Detection

Following Lau et al. (2012), to detect novel senses of a target word using this WSI method, we *jointly* topic model two corpora: a reference corpus — taken to represent standard usage — and a focus corpus of newer texts potentially containing novel senses. In other words, we extract usages of a target word w from both corpora, and then topic model the pooled instances of w. Under this approach, the discovered topics are applicable to both corpora, so there is no need to reconcile two different sets of topics. For the experiments in this paper, we extract three sentences of context for each usage, one sentence to either side of the usage of the target word.

As each usage is given a sense assignment, we can identify novel senses — senses present in the focus corpus, but unattested in the reference corpus — based on differences in the sense distribution for a given word between the two corpora. Lau et al. present a Novelty score which is proportional to the following:

$$\text{Novelty}_{\text{Ratio}}(s) = \frac{p_f(s)}{p_r(s)} \tag{1}$$

where $p_f(s)$ and $p_r(s)$ are the proportion of usages of a given word corresponding to sense s in the focus corpus and reference corpus, respectively, calculated using smoothed maximum likelihood estimates. The score for a given lemma is the maximum score for any of its induced senses. We refer to the *novel sense* for a lemma as the induced sense corresponding to this maximum.

4.2 Alternative Formulations of Novelty

The WSI system underlying the approach of Lau et al. labels each usage of a target lemma with an induced sense. Therefore, any approach to identifying keywords — words that are substantially more frequent in one corpus than another — can potentially be applied to identify novel senses, by viewing "words" as (word,sense) tuples. We consider a version of Novelty based on the difference in relative frequency of an induced sense in the focus and reference corpora, as below:

$$\text{Novelty}_{\text{Diff}}(s) = p_f(s) - p_r(s) \tag{2}$$

We consider a further new variant of Novelty based on the log-likelihood ratio of an induced sense in the two corpora, referred to as $\text{Novelty}_{\text{LLR}}$.

4.3 Incorporating knowledge of expected topics of novel senses

Cook et al. (2013) extended Lau et al.'s method by incorporating the observation that many neologisms are related to topics that are culturally salient (e.g., Ayto, 2006); nowadays we see many neologisms related to computing and the Internet. Indeed this observation was used to construct the gold-standard dataset for this study. Cook et al. identified a set of words, W, related to computing and the Internet, based on manual analysis of keywords for the corpora they considered. They then formulated the Relevance of an induced sense s for a given word as follows:

$$\text{Relevance}_{\text{Manual}}(s) = \sum_{w \in W} p(w|s) \tag{3}$$

For a given lemma, $\text{Relevance}_{\text{Manual}}$ is the maximum of this score for any of its induced senses, similar to Novelty.

Following Cook et al., we calculate Relevance and Novelty for each induced sense of each lemma, and then rank all the induced senses by these measures independently. We then compute the rank sum of each induced sense of each lemma under these two rankings. The final score for a given lemma is then the rank sum of its highest-ranked sense, and this sense is taken as that lemma's novel sense. We refer to this new method as "Rank Sum". Cook et al. only considered Novelty and Rank Sum; here we additionally consider Relevance on its own.

For the keywords, we manually construct a set of words related to computing and the Internet, the topics for which we expect to observe many novel senses in both of our datasets, in a similar way to Cook et al. In order to minimize annotation effort, we concentrate on words that are more-frequent in the

focus corpus than the reference corpus. For a given corpus pair, we begin by computing the keywords for those corpora using Kilgarriff's (2009) method.[8] Two annotators — both computational linguists and not authors of this paper — independently scanned the top-1000 keywords for the focus corpus, and selected those that were, based on their intuition, related to computing and the Internet. We then took the topically-relevant words for a given corpus pair to be those in the intersection of the sets of words selected by the two annotators. For BNC–ukWaC and SiBol/Port this gives 102 and 30 topically-relevant words, respectively. This annotation required, on average, 23 minutes per annotator per corpus pair to complete. Examples of the keywords selected for SiBol/Port include *broadband*, *click*, *device*, *online*, and *tweet*.

4.4 Automatically-extracting keywords

We propose a new fully-automated method for identifying a set of topically-relevant keywords. Because of the differences in corpus composition, the BNC–ukWaC keywords are often related to computing and the Internet. To automatically obtain topically-relevant words, we take the top-1000 keywords for the ukWaC relative to the BNC (i.e., the same keywords annotated for the BNC–ukWaC in Section 4.3). The keywords for SiBol/Port are less-clearly related to the topics of interest, so we therefore use the topically-relevant keywords from BNC–ukWaC for both datasets.

5 Results

In the following subsections we consider results at the type and then token level.

5.1 Type-level results

In these experiments we rank all items — lemmas with a novel sense, and distractors — by the various Novelty, Relevance and Rank Sum methods for the BNC–ukWaC and SiBol/Port datasets. 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 focus corpus and the reference corpus. This baseline has not been previously considered by Lau et al. (2012) or Cook et al. (2013).

To compare approaches, we examine precision–recall curves in Figures 1 and 2. In an applied setting, we envision these ranked lists being manually examined; we are therefore primarily interested in the highly-ranked items, i.e., the left portion of the precision–recall curves.

For BNC–ukWaC (Figure 1), $\text{Novelty}_{\text{Diff}}$ and $\text{Novelty}_{\text{Ratio}}$ perform much better than $\text{Novelty}_{\text{LLR}}$, but not better than the frequency ratio baseline, at least for the left-most portion of the precision–recall curve. Surprisingly, for Relevance, $\text{Relevance}_{\text{Auto}}$ outperforms $\text{Relevance}_{\text{Manual}}$. This could be because the focus corpus exhibits a clear topical bias towards computing and the Internet (the expected domain of many neologisms in the focus corpus), and therefore a larger set of potentially noisy keywords is more informative than a smaller, hand-selected set. All of the measures including the baseline, except for $\text{Novelty}_{\text{LLR}}$, assign higher scores to lemmas with a gold-standard novel sense than the distractors, according to a one-sided Wilcoxon rank sum test ($p < 0.05$ in each case).

Turning to SiBol/Port in Figure 2, the frequency ratio baseline is much less effective here; the frequency of the gold-standard novel senses is much lower overall than for BNC–ukWaC. All of the Novelty and Relevance methods outperform the baseline, and — with the exception of $\text{Novelty}_{\text{Ratio}}$ — rank the lemmas with a gold-standard novel sense higher than the distractors (again using a one-sided Wilcoxon rank sum test and $p < 0.05$). Furthermore, in this case, $\text{Relevance}_{\text{Manual}}$ outperforms $\text{Relevance}_{\text{Auto}}$, as expected.

In terms of the three Novelty measures, only $\text{Novelty}_{\text{Diff}}$ ranked items with a novel sense higher than the distractors for both datasets. We therefore also show results for the Rank Sum approach combining $\text{Novelty}_{\text{Diff}}$ and each of $\text{Relevance}_{\text{Manual}}$ and $\text{Relevance}_{\text{Auto}}$, denoted $\text{Rank Sum}_{\text{Diff,manual}}$ and $\text{Rank Sum}_{\text{Diff,auto}}$, respectively, in Figures 1 and 2. For both BNC–ukWaC and SiBol/Port, $\text{Rank Sum}_{\text{Diff,manual}}$

[8] Using this method, the keywordness score for a given word is simply the ratio of its frequency per million words, plus a constant, in two corpora; we set the constant to 100, the value recommended by Kilgarriff.

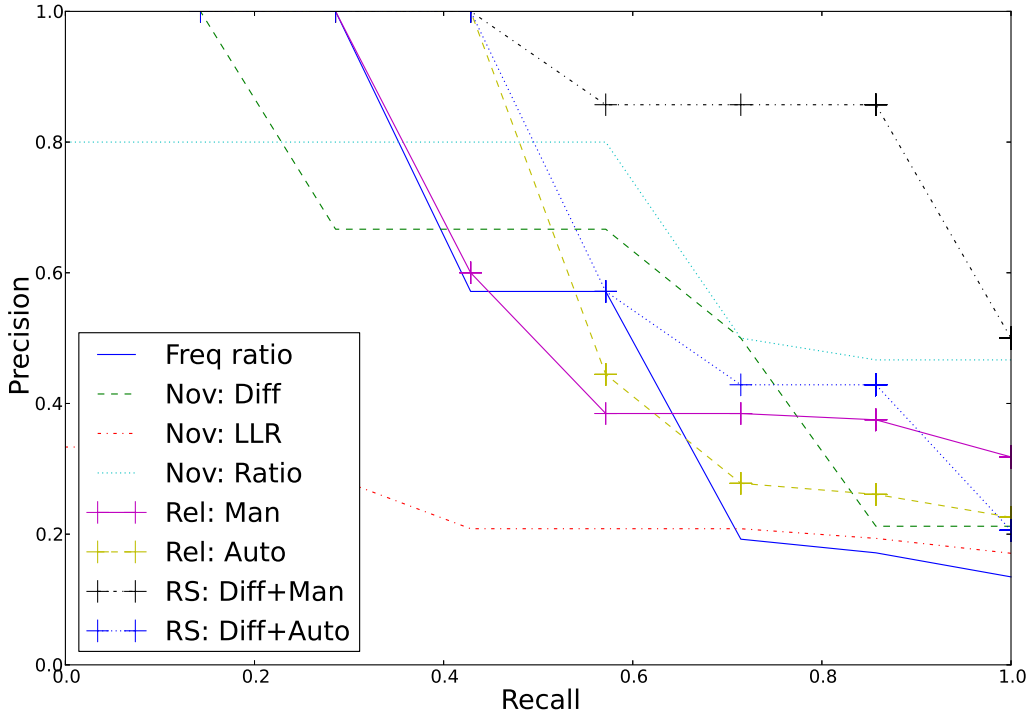

Figure 1: Precision–recall curve for the BNC–ukWaC dataset.

gives the best performance, and is a clear improvement over either of the individual methods. As expected, the performance of Rank $\text{Sum}_{\text{Diff,auto}}$ is not as good, but is nevertheless an improvement over the frequency ratio baseline for both datasets and provides an alternative to manual scrutiny of the keywords.

To further examine the potential of incorporating knowledge of the expected domains of novel senses to improve novel sense identification, we consider the case of *cloud* (n) from the SiBol/Port dataset. The highest-probability words for the topic with highest $\text{Novelty}_{\text{Diff}}$ are the following: *ash*, *volcanic*, *flight*, *@card@*,[9] *travel*, *airline*, *volcano*, *airport*, *air*, *cloud*. This sense appears to be related to the eruption of the Eyjafjallajökull volcano, a major event in 2010 (the year from which the SiBol/Port focus corpus is taken). Such topical differences, which do not correspond to a novel sense, are a problem for any approach to identifying lexical semantic differences between two corpora based on differences in the lexical context of a target word, and indeed observations such as this motivated our use of the methods incorporating Relevance. The highest probability words for the topic with highest $\text{Relevance}_{\text{Auto}}$ are the following: *cloud*, *@card@*, *company*, *service*, *business*, *computing*, *market*, *security*, *datum*, *need*. This topic appears to correspond to the expected novel sense of Internet-based computational resources, demonstrating the potential to improve a system for identifying novel word-senses by incorporating knowledge of the expected domains of neologisms. Moreover, incorporating Relevance is particularly powerful for avoiding false positives. For example, the distractor *clause* (n) is the lemma with the sixth-highest $\text{Novelty}_{\text{Diff}}$ for SiBol/Port. The highest probability words for the corresponding topic are the following: *contract*, *@card@*, *club*, *player*, *million*, *england*, *capello*, *manager*, *sign*, *deal*. This induced sense appears to be related to clauses in Fabio Capello's contract as manager of the England national football team, and is not a novel sense of *clause*. However, none of the induced senses of *clause* have high $\text{Relevance}_{\text{Auto}}$ or $\text{Relevance}_{\text{Manual}}$, and so incorporating information from Relevance can avoid incorrectly identifying this lemma as having a novel sense.

[9] A generic token signifying a cardinal number.

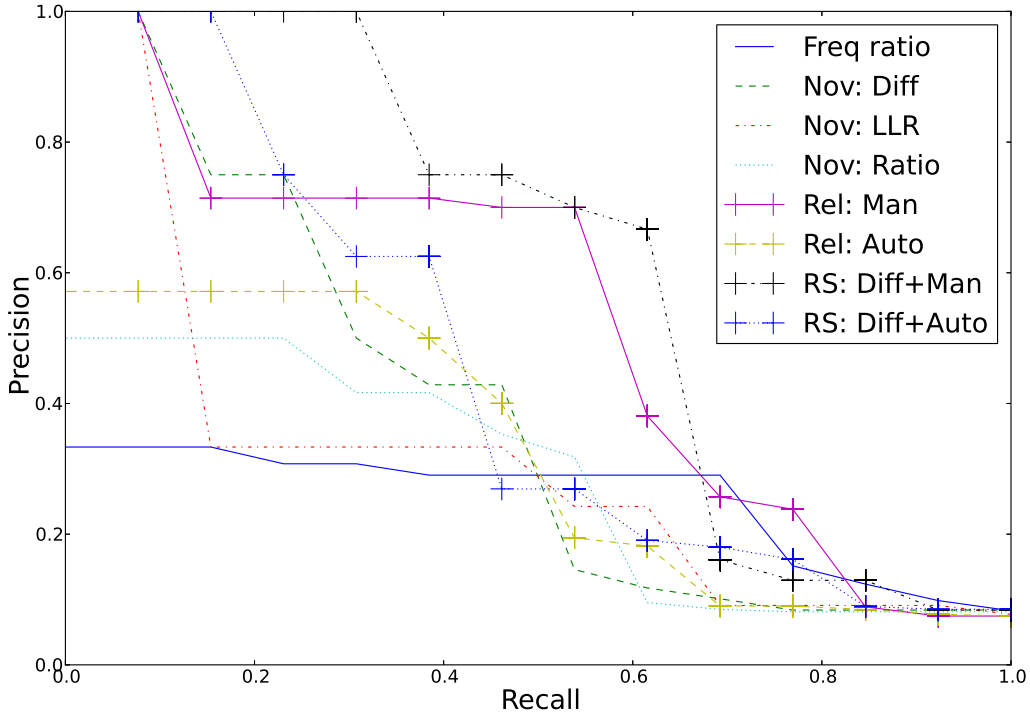

Figure 2: Precision–recall curve for the SiBol/Port dataset.

5.2 Token-level results

In this section, we consider the token-level identification of instances of the gold-standard novel senses. We compare Novelty, Relevance, and Rank Sum to a baseline that assigns all usages of a lemma to a single topic which is selected as the novel sense; in this case recall is 1, and precision is proportional to the frequency of the novel sense. We further consider the theoretical upper-bound of a method which selects a single topic as the novel sense, based on the output of the HDP-based WSI method; this oracle selects the best topic in terms of F-score as the novel sense. Results are presented in Table 2.

Each variant of Novelty and Relevance is an improvement over the baseline, although the Relevance measures don't perform as well as the Novelty ones, despite this dataset only containing novel senses related to computing (despite our efforts to include non-technical novel senses). For consistency with the presentation of the type-level results, we again consider Rank Sum using $\text{Novelty}_{\text{Diff}}$, even though it doesn't perform as well as $\text{Novelty}_{\text{LLR}}$ or $\text{Novelty}_{\text{Ratio}}$ on BNC–ukWaC. Using either automatically- or manually-obtained keywords, the performance of Rank Sum on BNC–ukWaC is remarkably on par with the upper-bound, although for SiBol/Port there is little or no improvement over $\text{Novelty}_{\text{Diff}}$. Nevertheless, these findings are further indication that novel sense identification can be improved by incorporating information about the topics for which we expect to see novel senses. However, this approach is particularly helpful at the type-level, where information about the expected topics of novel senses prevents lemmas not having a novel sense (i.e., the distractors) from being assigned high novelty.

6 Discussion and conclusion

The methods considered in this paper could be applied to any corpus pair, and potentially to identify lexical semantic differences between, for example, domains or language varieties. The focus of this study is English; sufficiently-large comparable corpora of national varieties of English (e.g., British and American English), are not readily-available, but could potentially be inexpensively constructed in the future (Cook and Hirst, 2012). We conducted some preliminary experiments using domain-specific sports

Method	F-score	
	BNC–ukWaC	SiBol/Port
$\text{Novelty}_{\text{Diff}}$	0.57	0.29
$\text{Novelty}_{\text{LLR}}$	0.67	0.28
$\text{Novelty}_{\text{Ratio}}$	0.66	0.28
$\text{Relevance}_{\text{Auto}}$	0.48	0.24
$\text{Relevance}_{\text{Manual}}$	0.45	0.27
$\text{Rank Sum}_{\text{Diff,auto}}$	0.72	0.30
$\text{Rank Sum}_{\text{Diff,manual}}$	0.72	0.29
Upper-bound	0.72	0.42
Baseline	0.36	0.20

Table 2: Token-level F-score for the BNC–ukWaC and SiBol/Port datasets using variants of Novelty, Relevance, and Rank Sum. The F-score of an oracle upper-bound and baseline are also shown.

and finance corpora (Koeling et al., 2005) and the BNC. However, in these experiments we observed very high $\text{Novelty}_{\text{Ratio}}$ for many distractors (selected in a similar way to our other experiments). Unlike the case of time difference, in corpora from different domains, an arbitrarily chosen word will tend to cooccur with very different words in the corpora, and $\text{Novelty}_{\text{Ratio}}$ will consequently be high. To address vocabulary differences between corpora, in their experiments on identifying lexical semantic differences between Dutch dialects, Peirsman et al. (2010) restricted the context words used to represent a target word to those with moderate frequency in each of the two corpora used. We considered a similar restriction in experiments on SiBol/Port, but did not see an overall improvement in performance.

We demonstrated that the performance of a method for identifying novel word-senses can be improved by incorporating information — acquired manually or automatically — about the expected topics of novel senses, which tend to be related to culturally-salient concepts. In future work, we intend to consider improved approaches for automatically identifying topically-relevant words by incorporating information about the top keywords of a corpus harvested from the Web for the domain of interest (e.g., PVS et al., 2012). We also believe that topic models could be useful for identifying emerging or changing domains themselves given the reference and focus corpus, and related work in this area (e.g., Wang and McCallum, 2006; Blei and Lafferty, 2007).

To conclude, we have presented the largest type- and token-level dataset of diachronic sense differences to date, drawing on two pairs of corpora, and have made this dataset available. We applied a recently-proposed WSI-based method to the task of finding sense differences in this data. We demonstrated that, while the method shows promise, on a type-based task it is comparable to a a simple frequency baseline, which had not been previously considered for this task. We carried out the first empirical evaluation of a recently-proposed extension of this method that incorporates manually-acquired knowledge of the expected domains of new senses, and found it to have superior performance at both the type and token level. We further proposed and evaluated an approach that only uses this domain knowledge, and a method for automating its acquisition.

Acknowledgments

We thank Michael Rundell and Macmillan Dictionaries for providing the list of headwords added to MEDAL since its first edition, and Charlotte Taylor for providing us with early access to SiBol/Port. We also thank Richard Fothergill, Karl Grieser, and Andrew Mackinlay for their help in annotation. This research was supported in part by funding from the Australian Research Council.

References

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.

David M. Blei and John D. Lafferty. 2007. Latent dirichlet allocation. *The Annals of Applied Statistics*, 1(1):17–35.

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.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

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.

Paul Cook and Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274. Singapore.

Paul Cook and Graeme Hirst. 2012. Do Web corpora from top-level domains represent national varieties of English? In *Actes des 11es Journées Internationales d'Analyse Statistique des Données Textuelles / Proceedings of the 11th International Conference on Textual Data Statistical Analysis*, pages 281–293. Liège, Belgium.

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 *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, pages 49–65. Tallinn, Estonia.

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.

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.

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.

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. 2009. Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress (EURALEX 2004)*, pages 105–116. Lorient, France.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of Human Language Technology Conference and Conference*

on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pages 419–426. Vancouver, Canada.

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.

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

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

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.

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

Avinesh PVS, Diana McCarthy, Dominic Glennon, and Jan Pomikálek. 2012. Domain specific corpora from the Web. In *Proceedings of the 15th Euralex International Congress*, pages 336–342. Oslo, Norway.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, pages 305–310. Portland, USA.

Michael Rundell and Gwyneth Fox, editors. 2002. *Macmillan English Dictionary for Advanced Learners*. Macmillan Education, Oxford, UK.

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.

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

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.

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

Xuerei Wang and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the Eleventh International Conference on Knowledge Discovery and Data Mining*, pages 424–433. Philadelphia, USA.

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.