Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples

Paul Cook,[1] Michael Rundell,[2] Jey Han Lau,[3] and Timothy Baldwin[1]
1. Department of Computing and Information Systems, The University of Melbourne
2. Lexicography Masterclass and Macmillan Dictionaries
3. Department of Philosophy, King's College London
E-mail: paulcook@unimelb.edu.au, michael.rundell@lexmasterclass.com, jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

There have been many recent efforts to automate or semi-automate parts of the process of compiling a dictionary, including building headword lists and identifying collocations. The result of these efforts has been both to make lexicographers' work more efficient, and to improve dictionaries by introducing more systematicity into the process of their construction. One task that has already been semi-automated is that of finding good dictionary examples, and a system for this, GDEX, is readily available in the Sketch Engine. An ideal system, however, would be able to automatically retrieve candidate examples of a particular *sense* of a word, which is beyond the current scope of GDEX. In this paper, as a step towards this ambitious goal, we propose and evaluate a method for applying a 'word-sense induction' system to automatically extract examples that exhibit a greater diversity of usages of a target word than can currently be obtained through GDEX. We then discuss the future prospects for systems that are able to automatically select candidate dictionary examples for a particular word sense.

Keywords: dictionary examples; word-sense induction; computational lexicography

1 Automating Lexicography

A major challenge facing contemporary lexicography – in the commercial sector, at least – is the goal of maximizing the potential of digital media and abundant language data, against a background of limited financial resources. As a response to this challenging climate, there has been significant progress in the automation of some of the tasks involved in compiling dictionaries and lexical databases – an approach which has the potential to deliver reduced development costs together with improved coverage of lexicographically-relevant facts (e.g., Rundell & Kilgarriff 2011, Cook et al. 2013, Kosem, Gantar & Krek 2013).

One component of this project is the automatic retrieval from corpus data of 'good' dictionary examples, whether for presenting editors with a shortlist of appropriate candidates, populating a dictionary database with examples, or providing the end-user with a range of instances of words in context (Kilgarriff et al. 2008; Kosem, Husák & McCarthy 2011). Notwithstanding these advances, there is scope for improvement in two areas. First, example-finding software does not yet routinely achieve the contextual diversity that characterizes example-sets selected by skilled lexicographers. Secondly, it does not attempt the difficult but critical task of mapping corpus instances onto dictionary senses. Some current dictionaries provide a range of (automatically-retrieved) examples to complement the manually-selected ones in the dictionary. This approach can be found, for example, in the 5th edition of the *Longman Dictionary of Contemporary English* (*LDOCE*) (http://ldoce.longmandictionariesonline.com/dict/SearchEntry.html), where users can opt to see up to ten 'Examples from the corpus', and in *Wordnik* (www.wordnik.com), where numerous authentic examples are shown on the right-hand side of the screen. Google Translate (http://translate.google.com) has a somewhat similar feature. But in none of these cases are the examples attached to word senses: in *LDOCE*, the entry for *pool* (noun) includes a random assortment of examples, including references to swimming pools, pools of investment funds, and even football pools. Ideally, we need a system which automatically selects optimally-diverse examples for a polysemous word (so that users are offered examples exhibiting the full contextual range of the word's behaviour) and matches the examples to the individual dictionary senses whose meaning they instantiate.

In this paper we report an experiment in which a 'word sense induction' methodology is applied to extracting corpus examples in a way that fulfills the first of these goals – identifying examples showing the diversity of contexts in which a word is used. We conclude by discussing the prospects for using the output of the word-sense induction system to map the 'induced senses' the system discovers in the corpus to dictionary senses.

2 Diversifying Example Sentences with Word-sense Induction

In this section we describe the GDEX method for identifying good dictionary examples and a recently-presented word-sense induction (WSI) system, and then propose a method to combine these two technologies to automatically select more-diverse dictionary examples.

Cook, Paul, Michael Rundell, Jey Han Lau and Timothy Baldwin (to appear) Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples, In <em>Proceedings of the XVI EURALEX International Congress (EURALEX 2014), Bolzano, Italy, pp. 15-19.

2.1 GDEX

GDEX (Kilgarriff et al. 2008) is a system for automatically selecting good dictionary examples from a corpus. Sentences containing a given target word are scored based on a number of heuristics about what makes a sentence a good dictionary example, such as sentence length, the position of the target word in the sentence, and the other words occurring in the sentence. Given a query for a particular word, GDEX then returns the top-scoring sentences in a corpus, which can be manually examined for selection as examples. GDEX has become a standard lexicographical tool, and is available for use with many corpora in the Sketch Engine (SkE, http://www.sketchengine.co.uk/, Kilgarriff & Tugwell 2002). Adaptations of GDEX (Kosem, Gantar & Krek 2011) have incorporated simple notions of diversity to avoid selecting duplicate or very similar sentences. We propose a more sophisticated notion of diversity targeted at selecting a set of candidate example sentences exhibiting a wider range of senses of the target word.

2.2 Word-sense Induction

WSI is 'the task of automatically grouping the usages of a given word in a corpus according to sense, such that all usages exhibiting a particular sense are in the same group' (Cook et al. 2013). Crucially this grouping is done without reference to a pre-existing sense inventory. WSI is the automatic counterpart to the manual lexicographic process of word sense disambiguation (WSD, Atkins & Rundell 2008: 269).[1] Although the notion of word sense is of course controversial, it is nevertheless standard for dictionaries to carve up the meanings of a word into senses, even though dictionaries will vary in terms of the sense distinctions made for a given polysemous word.

Topic modeling (Blei, Ng & Jordan 2003) is a computational technique for automatically discovering latent structure in a corpus that has recently been successfully applied to a wide range of NLP tasks. A typical topic model automatically 'learns' the topics in a corpus, and the mixture of topics in each document in the corpus. Each topic is represented as a probability distribution over words; each document is represented as a probability distribution over topics.

Lau et al. (2012) present a WSI system based on topic modeling. Rather than building a topic model for an entire corpus, they build a separate model for each target word. In this model the "documents" are short contexts – typically 3 sentences – containing a usage of the target word. There is not necessarily a correspondence between the topics in a topic model and topics in the sense of the subject of a text (although a topic model for a corpus will often learn topics that do indeed correspond to this more common usage of *topic*). In Lau et al.'s WSI methodology, the topics in the topic model are interpreted as word senses.

In traditional topic models (e.g., latent Dirichlet allocation, Blei, Ng & Jordan 2003) the number of topics to be learned must be specified manually in advance. For WSI this would mean that the number of senses for each word would need to be set by hand. Words of course differ with respect to their polysemy, and an appropriate number of senses could only be determined based on corpus analysis. Lau et al. therefore use hierarchical Dirichlet process (Teh et al. 2006), a type of topic model that also automatically learns the appropriate number of topics for a given document collection.

In Lau et al.'s model, each 'document' – typically consisting of a sentence including an instance of the target word, and one sentence of context before and after – is represented as a bag-of-words, i.e., word order is ignored, but the frequency of words in the context is maintained. Because the immediate context surrounding a word can be highly informative of that word's sense, positional word features that encode the specific three words occurring to the left and right of the target word are also included. In this document representation stopwords are ignored, and all other words are lemmatized. An example of the representation used is given in Table 1. For reasons of brevity the 'document' in this example is a single sentence (whereas it would typically be three sentences). An example of the senses induced by the model for the lemma *box* (n) is given in Table 2. Recall that each sense is a probability distribution over words; here the top-10 most likely terms for each sense are shown. Examining these terms allows us to roughly interpret the senses the system has induced. For example, senses 1 and 2 seem to correspond to usages of *box* in the context of sports and elections, respectively. Random samples of 5 corpus instances corresponding to induced senses 1 and 4 for *box* (n) are given in Table 3. For sense 1, all of the usages seem to correspond to an area of a sports ground, although the first four relate to soccer, but the last relates to tennis. For sense 4, the model has identified usages related to the entertainment industry, but includes a mixture of usages of the expressions *box office* and *box set*, as well as other usages such as the final example.

[1] It is worth clarifying a terminological difference between the lexicographic and natural language processing (NLP) communities. In NLP, WSD refers specifically to the task of selecting the most appropriate sense, from a given sense inventory, for a given instance of a word in context. In lexicography WSD typically refers to identifying the various senses of a word, i.e., constructing a sense inventory, based on corpus evidence.

Target Lemma	box (n)
‘Document’	The Flames had a two-man advantage near the end of the second period when Lacroix and McSorley were in the penalty **box** for kneeing and unsportsmanlike conduct, respectively.
Bag-of-words Features	flame, two-man, advantage, near, end, second, period, lacroix, mcsorley, penalty, knee, unsportsmanlike, conduct, respectively
Positional Word Features	lacroix_#-3, mcsorley_#-2, penalty_#-1, knee_#+1, unsportsmanlike_#+2, conduct_#+3

Table 1: An example of the topic model features.

Sense Number	Top-10 Terms
1	box minute @card@ game ball goal score shot penalty play
2	box @card@ ballot ballot_#-1 police official election vote find party
3	box @card@ company computer digital cable converter black_#-1 black converter_#-1
4	office box office_#+1 million @card@ film movie weekend dollar ticket
5	box @card@ n’t look find small think put room store
6	box cereal recipe cut outlet post fat gram vip wire
7	shanker teller lionel penn crush ferry jesus julie maine marshal

Table 2: The top-10 terms for each of the senses induced for the lemma *box* (n).

Sense Number	Usage
1	Bolivia added an extra insurance goal in the 80th when Ronaldo Garcia sent a long blast from outside the **box** into the upper corner. Arsenal were left nursing a justifiable grievance over the referee's failure to award a second-half penalty when Kuyt unbalanced Alexander Hleb with a tug from behind as the midfielder wriggled into space deep in the **box**. Thiago made it 3-0 in the 79th minute with a powerful left foot drive from the edge of the **box**. Masami Ihari headed away a corner from Juninho but only as far as Zinho, who let fly with a spectacular volley from the edge of the **box** that gave no chance at all. He's placing it in the **box** beautifully, hitting closer to the line with lots of aces.
4	And while Joel is paying himself a backhanded compliment, especially in the context of B-sides, live tracks and rarities **box** set, it got me thinking. It's the harsh command of the **box** office, demanding a big seller and the heck with all else. "For the most part, it'll help us," says Bobbie Welch, **box** office manager for New Mexico State University, which has hosted ZZ Top, Guns N' Roses and Paul McCartney. For those seeking more kid-friendly fare, the "Spotlight Collection" discs – including a sixth released Tuesday ($27) – offer more than 30 family-appropriate installments from the Golden **box** sets. Only 17 percent of reviews were positive, according to RottenTomatoes.com, but 82 percent of audience survey respondents checked off the "excellent" or "very good" **boxes**, according to Sony.

Table 3: Usages corresponding to induced senses 1 and 4 of the lemma *box* (n).

Lau, Cook & Baldwin (2013a,b) recently showed this WSI methodology to be the overall best performing system on two recent SemEval WSI shared tasks (Jurgens & Klapaftis 2013; Navigli & Vannella 2013). Cook et al. (2013) demonstrated that this system can be applied as a lexicographical tool for finding new word-senses. We therefore adopt this WSI system here for diversifying automatically-selected dictionary examples.

2.3 Diversification

For a given target lemma, we obtain the top-100 GDEX examples for a corpus from SkE. We further obtain a random sample of up to 50k usages of the target from the same corpus. In each case we extract the sentence containing the usage of the target, and one sentence of context on either side. Following Lau et al. (2012) we remove stopwords and lemmatize the tokens in the context. We then run the WSI system on these usages of the target lemma. The WSI system outputs a label indicating the induced sense number of each target instance. These induced senses correspond to groups of usages that exhibit the same sense, according to the WSI system, not dictionary senses.

We then use these induced sense labels to diversify the top-100 GDEX examples. To do so, we repeatedly iterate through the top-100 GDEX sentences (i.e., we consider each sentence in turn, one by one). For each pass over the sentences, we select the best GDEX sentence (according to GDEX's ranking) for each induced sense, which has not been selected in a previous pass. We repeat this until all sentences have been selected. In the subsequent analysis we compare the top-5 GDEX examples to the top-5 examples produced by this diversification procedure.

3 Analysis

For this preliminary analysis we selected 98 target lemmas to analyse: 54 from a recent SemEval WSI task (Jurgens & Klapaftis 2013) and 44 additional medium-polysemy lemmas. We extracted GDEX sentences, and the additional randomly-selected usages, from the ukWaC (Ferraresi et al. 2008).

We ran the GDEX diversification procedure described in the previous section for each target lemma. The top-5 GDEX sentences for the target lemma exhibited varying numbers of induced senses (as determined by the WSI software). However, if the top-5 GDEX sentences already exhibit many (e.g., 4 or 5) induced senses, then our diversification procedure has little or no impact. We therefore focused our analysis on lemmas where the top-5 GDEX usages exhibited less diversity. Crucially such cases can easily be automatically identified. Here we discuss the findings for twelve lemmas whose top-5 GDEX sentences were the least diverse, exhibiting just two induced senses of the target lemma in each case. (In no case did the top-5 GDEX usages exhibit just one induced sense.)

For each lemma, two sets of five example sentences were prepared: (1) the top-5 GDEX sentences, and (2) the top-5 sentences from our diversification procedure. These sets of sentences were presented to a professional lexicographer (the second author of this paper) who was asked to judge which set of examples was better. Crucially the lexicographer did not know which method the sets of examples corresponded to. For eight lemmas the examples produced through our new diversification procedure were selected as better; in the remaining four cases the default GDEX examples were chosen. To give an idea of the potential of our method, we discuss the output of the two systems for two of the target lemmas which were analysed: *exploitation* and *bitter*.

Top-5 sentences from GDEX:

1. It must be a world where humanity has been liberated from social distress, brutal **exploitation** and war.
2. A new minister replaces the old one, the daily grind of **exploitation** resumes.
3. And Engage, wittingly or not, is aiding it in this **exploitation**.
4. It's not trade we're against, it's **exploitation** and unchecked power.
5. Others have seen Napster as little more than payback for decades of record company **exploitation** of artists.

Top-5 sentences from our new diversification approach:

1. It must be a world where humanity has been liberated from social distress, brutal **exploitation** and war.
2. Others have seen Napster as little more than payback for decades of record company **exploitation** of artists.
3. Delivery will focus upon the development and **exploitation** of repertoire through the workplace.
4. Requirement 24 - Identify all auditable events that may be used in exploitation of known covert storage channels.
5. Workers can refer young people they think are vulnerable or at risk of homelessness and sexual **exploitation**.

The examples that are shared by the two sets are both good. However, the second and third sentences from default GDEX are weak because they offer little context for interpretation and include anaphora. Although the final sentence for default GDEX (sentence 4) is acceptable, the diversified sentences provide a better snapshot of the word overall, in that they cover the more neutral sense of exploitation (sentence 4) and include an example of *sexual exploitation* (sentence 5).

Top-5 sentences from GDEX:

1. If cows eat too many carrots, their milk tastes **bitter**.
2. It’s so easy for us to be **bitter**.
3. Men obeyed their base immediate motives until the world grew unendurably **bitter**.
4. In his destroyed Urim, its lament is **bitter**.
5. No; and henceforth I can never trust his word. **Bitter**, bitter confession!

Top-5 sentences from our new diversification approach:

1. If cows eat too many carrots, their milk tastes **bitter**.
2. It’s so easy for us to be **bitter**.
3. In his destroyed Urim, its lament is **bitter**.
4. The rivalry between Soka Gakkei and Aum Shinrikyo was **bitter**.
5. Joe Frazier never felt more **bitter** about defeat, and continues even today to hate his great rival.

In this case, three of the examples appear in both sets, and their quality varies: the first in each set is good; the second acceptable, if a little short on context; the third (*In his destroyed Urim…*) is weak, and would certainly not make an appropriate example sentence for a pedagogical dictionary (or any other dictionary, for that matter). But if we compare the two examples unique to each set, those in the second set (sentences 4 and 5) are clearly more suitable as dictionary examples than those in the first (3 and 5). *Rivalry* has a high saliency score as a collocate of *bitter* (though *disappointment* would have been even better, assuming the goal is maximum typicality). And the addition of a collocating preposition in the last example (with *about*, the most frequent preposition appearing with *bitter*) provides additional diversity.

4 Discussion

Software for selecting dictionary examples could be improved if it were optimised to select diverse examples for a polysemous word, such that the examples show the full range of usage for that word. In this paper we have proposed a novel method for automatically selecting a more diverse set of dictionary examples from a corpus than can currently be obtained using GDEX. We carried out a small-scale preliminary evaluation of this method, and found that – in terms of diversity – our approach outperformed GDEX for eight out of twelve lemmas analyzed. The results are encouraging rather than conclusive. But with further improvements based on what we have learned through this experiment, this new method could be applied to real lexicographic tasks – either for providing editors with candidate lists from which to select examples for a dictionary, or for automatically providing dictionary users with additional examples. In either case, the outcome should be a set of examples exhibiting a more diverse range of usages than current software tools usually supply.

Systems for selecting examples would be further improved if they were able to match automatically-identified examples to corresponding dictionary senses. Lau et al. (2014) recently proposed a method to link the senses induced by the same WSI system used here to senses in a dictionary. They evaluated this method in the context of identifying the relative frequencies of the senses of a given word in a corpus, and showed it to perform comparably to previously-proposed approaches for this task (McCarthy et al. 2007).[2] In future work, we intend to combine this method for linking induced senses to dictionary senses with the approach described in this paper, in order to identify good examples at the level of *word senses*, as opposed to *lemmas*. WSI remains a very difficult task for current natural language processing technologies. Crucially, in the context of identifying sense-specific dictionary examples, it might not be necessary to correctly identify the dictionary sense of every corpus instance of a target word. Instead, it might suffice to identify the dictionary sense corresponding to those instances where the system is highly confident of its prediction, and to then apply GDEX to select good dictionary examples amongst those instances. We are therefore optimistic about the future possibility of automatically adding sense specific examples to dictionaries, although much work remains to be done.

The WSI system of Lau et al. (2012) lies at the core of the method presented in this paper. To encourage further research on WSI and its applications, Lau, Cook & Baldwin (2013a,b) made this system publicly available under a license which permits its use for commercial purposes (https://github.com/jhlau/hdp-wsi). We hope that others will make use of this software to consider further applications of topic modeling and WSI in computational lexicography.

5 References

Atkins, B. T. S. and Rundell, R. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cook, P., Lau, J.H., Rundell, M, McCarthy, D. & Baldwin, T. (2013). 'A lexicographic appraisal of an automatic approach for detecting new word-senses', in Kosem et al. 2013: 49-65.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, pages 47-54, Marrakech, Morocco.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290-299, Atlanta, USA.

Kilgarriff, A. and Tugwell, D. (2002). 'Sketching words'. In Corréard, M.-H., editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125-137. Euralex, Grenoble, France.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008) 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus', in Bernal, E. and DeCesaris, J. (Eds) Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra: 425-433.

Kosem, I., Gantar, P., & Krek, S. (2013). 'Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing', in Kosem et al. 2013: 32-48.

Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) (2013). *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Kosem, I., Husák, M., & McCarthy, D. (2011). 'GDEX for Slovene'. In I. Kosem, K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies: 151-159.

Lau, J. H., Cook, P., and Baldwin, T. (2013a). unimelb: Topic modelling-based word sense induction. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307–311, Atlanta, USA.

Lau, J. H., Cook, P., and Baldwin, T. (2013b). unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of*

[2] Very interestingly, but of less relevance to the present paper, they also showed that this method has the potential to identify dictionary senses that are unattested in a corpus, and senses that are induced by the WSI system but not listed in a dictionary.

the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 217–221, Atlanta, USA.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (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.

Lau, J. H., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA.

McCarthy, D., Koeling, R., Weeds, J., and Carroll J., (2007). Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics,* 33(4):553–590.

Navigli, R. and Vannella, D. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, USA.

Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In Meunier F., De Cock S., Gilquin G. and Paquot M. (Eds), *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Benjamins: 257-281.

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

Acknowledgements

This work was supported in part by the Australian Research Council.