Generating Example Contexts to Illustrate a Target Word Sense

Jack Mostow
Carnegie Mellon University
RI-NSH 4103, 5000 Forbes Avenue
Pittsburgh, PA 15213-3890, USA
mostow@cs.cmu.edu

Weisi Duan
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA 15213-3890, USA
wduan@cs.cmu.edu

Abstract

Learning a vocabulary word requires seeing it in multiple informative contexts. We describe a system to generate such contexts for a given word sense. Rather than attempt to do word sense disambiguation on example contexts already generated or selected from a corpus, we compile information about the word sense into the context generation process. To evaluate the sense-appropriateness of the generated contexts compared to WordNet examples, three human judges chose which word sense(s) fit each example, blind to its source and intended sense. On average, one judge rated the generated examples as sense-appropriate, compared to two judges for the WordNet examples. Although the system's precision was only half of WordNet's, its recall was actually higher than WordNet's, thanks to covering many senses for which WordNet lacks examples.

1 Introduction

Learning word meaning from example contexts is an important aspect of vocabulary learning. Contexts give clues to semantics but also convey many other lexical aspects, such as parts of speech, morphology, and pragmatics, which help enrich a person's word knowledge base (Jenkins 1984; Nagy *et al.* 1985; Schatz 1986; Herman *et al.* 1987; Nagy *et al.* 1987; Schwanenflugel *et al.* 1997; Kuhn and Stahl 1998; Fukkink *et al.* 2001). Accordingly, one key issue in vocabulary instruction is how to find or create good example contexts to help children learn a particular sense of a word. Hand-vetting automatically generated contexts can be easier than hand-crafting them from scratch (Mitkov *et al.* 2006; Liu *et al.* 2009).

This paper describes what we believe is the first system to generate example contexts for a given target sense of a polysemous word. Liu et al. (2009) characterized good contexts for helping children learn vocabulary and generated them for a target part of speech, but not a given word sense. Pino and Eskenazi (2009) addressed the polysemy issue, but in a system for selecting contexts rather than for generating them. Generation can supply more contexts for a given purpose, e.g. teaching children, than WordNet or a fixed corpus contains.

Section 2 describes a method to generate sense-targeted contexts. Section 3 compares them to WordNet examples. Section 4 concludes.

2 Approach

An obvious way to generate sense-targeted contexts is to generate contexts containing the target word, and use Word Sense Disambiguation (WSD) to select the ones that use the target word sense. However, without taking the target word sense into account, the generation process may not output any contexts that use it. Instead, we model word senses as topics and incorporate their *sense indicators* into the generation process – words that imply a unique word sense when they co-occur with a target word.

For example, *retreat* can mean "a place of privacy; a place affording peace and quiet." Indicators for this sense, in decreasing order of Pr(word | topic for target sense), include *retreat, yoga, place, retreats, day, home, center, church, spiritual, life, city, time, lake, year, room, prayer, years, school, dog, park, beautiful, area,* and *stay.* Generated contexts include …*retreat in this bustling* ***city***….

Another sense of *retreat* (as defined in WordNet) is "(military) a signal to begin a withdrawal

Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 105–110,
Portland, Oregon, 24 June 2011. ©2011 Association for Computational Linguistics

from a dangerous position," for which indicators include *states, war, united, american, military, flag, president, world, bush, state, Israel, Iraq, international, national, policy, forces, foreign, nation, administration, power, security, iran, force,* and *Russia*. Generated contexts include …***military*** *leaders believe that retreat*….

We decompose our approach into two phases, summarized in Figure 1. Section 2.1 describes the Sense Indicator Extraction phase, which obtains indicators for each WordNet synset of the target word. Section 2.2 describes the Context Generation phase, which generates contexts that contain the target word and indicators for the target sense.

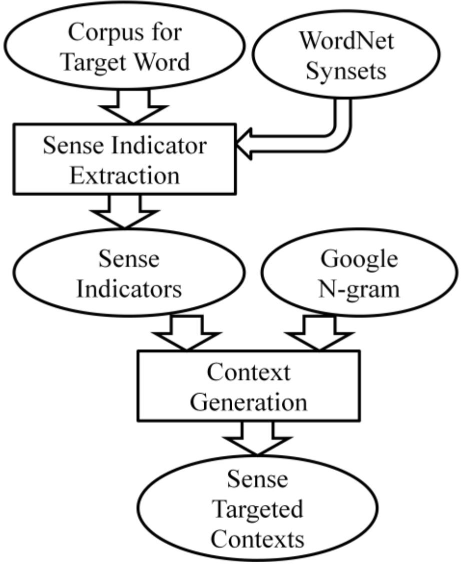

Figure 1: overall work flow diagram

2.1 Sense Indicator Extraction

Kulkarni and Pedersen (2005) and Duan and Yates (2010) performed Sense Indicator Extraction, but the indicators they extracted are not sense targeted. Content words in the definition and examples for each sense are often good indicators for that sense, but we found that on their own they did poorly.

One reason is that such indicators sometimes cooccur with a different sense. But the main reason is that there are so few of them that the word sense often appears without any of them. Thus we need more (and if possible better) sense indicators.

To obtain sense-targeted indicators for a target word, we first assemble a corpus by issuing a Google query for each synset of the target word. The query lists the target word and all content words in the synset's WordNet definition and examples, and specifies a limit of 200 hits. The resulting corpus contains a few hundred documents.

To extract sense indicators from the corpus for a word, we adapt Latent Dirichlet Allocation (LDA) (Blei *et al.* 2003). LDA takes as input a corpus of documents and an integer k, and outputs k latent topics, each represented as a probability distribution over the corpus vocabulary. For k, we use the number of word senses. To bias LDA to learn topics corresponding to the word senses, we use the content words in their WordNet definitions and examples as seed words.

After learning these topics and filtering out stop words, we pick the 30 highest-probability words for each topic as indicators for the corresponding word sense, filtering out any words that also indicate other senses. We create a corpus for each target word and run LDA on it.

Having outlined the extraction process, we now explain in more detail how we learn the topics; the mathematically faint-hearted may skip to Section 2.2. Formally, given corpus C with m documents, let n be the number of topics, and let α_i and β_j be the parameters of the document and topic distributions respectively. LDA assumes this generative process for each document D_i for a corpus C:

1. Choose $\theta_i \sim Dir(\alpha_i)$ where $i \in \{1, \dots, m\}$
2. Choose $\gamma_j \sim Dir(\beta_j)$ where $j \in \{1, \dots, n\}$
3. For each word $w_{i,t}$ in D_i where $t \in \{1, \dots, T\}$, T is the number of words in D_i
   (a) Choose a topic $z_{i,t} \sim Multinomial(\theta_i)$
   (b) Choose a topic $w_{i,t} \sim Multinomial(\gamma_s)$ where $s = z_{i,t}$

In classical LDA, all α_i's are the same. We allow them to be different in order to use the seed words as high confidence indicators of target senses to bias the hyper-parameters of their document distributions.

For inference, we use Gibbs Sampling (Steyvers and Griffiths 2006) with transition probability

$$P\left(z_{i,t} = z \middle| Z_{c \setminus z_{i,t}}, C, w_{i.t} = w\right) = \frac{count(w,z) + \beta_{z,w}}{count(z) + \sum_{\boldsymbol{W}} \beta_{z,\boldsymbol{W}}} * \frac{count_i(z) + \alpha_{i,z}}{\sum_{\boldsymbol{Z}} count_i(\boldsymbol{Z}) + \sum_{\boldsymbol{Z}} \alpha_{i,\boldsymbol{Z}}}$$

Here $Z_{c \setminus z_{i,t}}$denotes the topic assignments to all other words in the corpus except $w_{i,t}$; $count(w,z)$ is the number of times word w is assigned to topic z in the whole corpus; $count(z)$ is the number of words assigned to topic z in the entire corpus;

$count_i(z)$ is the count of tokens assigned to topic z in document D_i; and $\beta_{z,w}$and $\alpha_{i,z}$ are the hyper-parameters on $\gamma_{z,w}$ and $\theta_{i,z}$ respectively in the two Dirichlet distributions.

For each document D_i that contains seed words of some synset, we bias α_i toward the topic z for that synset by making $\alpha_{i,z}$ larger; specifically, we set each $\alpha_{i,z}$ to 10 times the average value of α_i. This bias causes more words w_{new} in D_i to be assigned to topic z because the words of D_i are likely to be relevant to z. These assignments then influence the topic distribution of z so as to make w_{new} likelier to be assigned to z in any document D_{new}, and thus shift the document distribution in D_{new} towards z. By this time we are back to the start of the loop where the document distribution of D_{new} is biased to z. Thus this procedure can discover more sense indicators for each sense.

Our method is a variant of Labeled LDA (L-LDA) (Ramage 2009), which allows only labels for each document as topics. In contrast, our variant allows all topics for each document, because it may use more than one sense of the target word. Allowing other senses provides additional flexibility to discover appropriate sense indicators.

The LDA method we use to obtain sense indicators fits naturally into the framework of bootstrapping WSD (Yarowsky 1995; Mihalcea 2002; Martinez *et al.* 2008; Duan and Yates 2010), in which seeds are given for each target word, and the goal is to disambiguate the target word by bootstrapping good sense indicators that can identify the sense. In contrast to WSD, our goal is to generate contexts for each sense of the target word.

2.2 Context Generation

To generate sense-targeted contexts, we extend the VEGEMATIC context generation system (Liu *et al.* 2009). VEGEMATIC generates contexts for a given target word using the Google N-gram corpus. Starting with a 5-gram that contains the target word, VEGEMATIC extends it by concatenating additional 5-grams that overlap by 4 words on the left or right.

To satisfy various constraints on good contexts for learning the meaning of a word, VEGEMATIC uses various heuristic filters. For example, to generate contexts likely to be informative about the word meaning, VEGEMATIC prefers 5-grams that contain words related to the target word, i.e., that occur more often in its presence. However, this criterion is not specific to a particular target sense.

To make VEGEMATIC sense-targeted, we modify this heuristic to prefer 5-grams that contain sense indicators. We assign the generated contexts to the senses whose sense indicators they contain. We discard contexts that contain sense indicators for more than one sense.

3 Experiments and Evaluation

To evaluate our method, we picked 8 target words from a list of polysemous vocabulary words used in many domains and hence important for children to learn (Beck *et al.* 2002). Four of them are nouns: *advantage* (with 3 synsets), *content* (7), *force* (10), and *retreat* (7). Four are verbs: *dash* (6), *decline* (7), *direct* (13), and *reduce* (20). Some of these words can have other parts of speech, but we exclude those senses, leaving 73 senses in total.

We use their definitions from WordNet because it is a widely used, comprehensive sense inventory. Some alternative sense inventories might be unsuitable. For instance, children's dictionaries may lack WordNet's rare senses or hypernym relations.

We generated contexts for these 73 word senses as described in Section 2, typically 3 examples for each word sense. To reduce the evaluation burden on our human judges, we chose just one context for each word sense, and for words with more than 10 senses we chose a random sample of them. To avoid unconscious bias, we chose random contexts rather than the best ones, which a human would likelier pick if vetting the generated contexts by hand. For comparison, we also evaluated WordNet examples (23 in total) where available.

We gave three native English-speaking college-educated judges the examples to evaluate independently, blind to their intended sense. They filled in a table for each target word. The left column listed the examples (both generated and WordNet) in random order, one per row. The top row gave the WordNet definition of each synset, one per column. Judges were told: ***For each example, put a 1 in the column for the sense that best fits how the example uses the target word. If more than one sense fits, rank them 1, 2, etc. Use the last two columns only to say that none of the senses fit, or you can't tell, and why.*** (Only 10 such cases arose.)

We measured inter-rater reliability at two levels.

At the fine-grained level, we measured how well the judges agreed on which one sense fit the example best. The value of Fleiss' Kappa (Shrout and Fleiss 1979) was 42%, considered moderate. At the coarse-grained level, we measured how well judges agreed on which sense(s) fit at all. Here Fleiss' Kappa was 48%, also considered moderate.

We evaluated the examples on three criteria.

Yield is the percentage of intended senses for which we generate at least one example – whether it fits or not. For the 73 synsets, this percentage is 92%. Moreover, we typically generate 3 examples for a word sense. In comparison, only 34% of the synsets have even a single example in WordNet.

(Fine-grained) ***precision*** is the percentage of examples that the intended sense fits best according to the judges. Human judges often disagree, so we prorate this percentage by the percentage of judges who chose the intended sense as the best fit. The result is algebraically equivalent to computing precision separately according to each judge, and then averaging the results. Precision for generated examples was 36% for those 23 synsets and 27% for all 67 synsets with generated examples. Although we expected WordNet to be a gold standard, its precision for the 23 synsets having examples was 52% — far less than 100%.

This low precision suggests that the WordNet contexts to illustrate different senses were often not informative enough for the judges to distinguish them from all the other senses. For example, the WordNet example *reduce one's standard of living* is attached to the sense "lessen and make more modest." However, this sense is hard to distinguish from "lower in grade or rank or force somebody into an undignified situation." In fact, two judges did not choose the first sense, and one of them chose the second sense as the best fit.

Coarse-grained precision is similar, but based on how often the intended sense fits the example at all, whether or not it fits best. Coarse-grained precision was 67% for the 23 WordNet examples, 40% for the examples generated for those 23 synsets, and 33% for all 67 generated examples.

Coarse-grained precision is important because fine-grained semantic distinctions do not matter in illustrating a core sense of a word. The problem of how to cluster fine-grained senses into coarse senses is hard, especially if consensus is required (Navigli *et al*. 2007). Rather than attempt to identify a single definitive partition of a target word's synsets into coarse senses, we implicitly define a coarse sense as the subset of synsets rated by a judge as fitting a given example. Thus the clustering into coarse senses is not only judge-specific but example-specific: different, possibly overlapping sets of synsets may fit different examples.

Recall is the percentage of synsets that fit their generated examples. Algebraically it is the product of precision and yield. Fine-grained recall was 25% for the generated examples, compared to only 18% for the WordNet examples. Coarse-grained recall was 30% for the generated examples, compared to 23% for the WordNet examples.

Figure 2 shows how yield, inter-rater agreement, and coarse and fine precision for the 8 target words vary with their number of synsets. With so few words, this analysis is suggestive, not conclusive. We plot all four metrics on the same [0,1] scale to save space, but only the last two metrics have directly comparable values, However, it is still meaningful to compare how they vary. Precision and inter-rater reliability generally appear to decrease with the number of senses. As polysemy increases, the judges have more ways to disagree with each other and with our program. Yield is mostly high, but might be lower for words with many senses, due to deficient document corpora for rare senses.

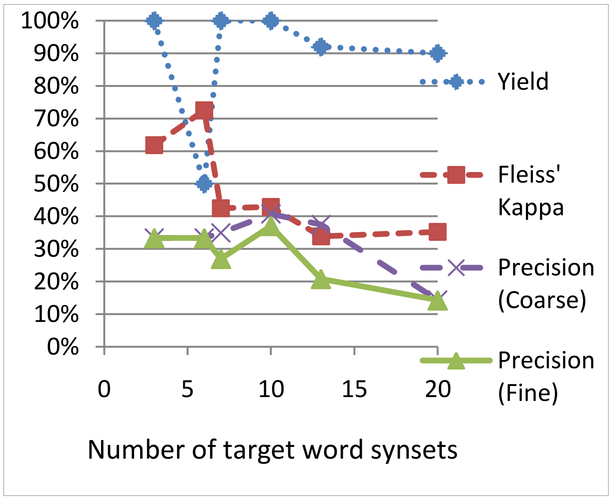

Figure 2: Effects of increasing polysemy

Table 1 compares the generated and WordNet examples on various measures. It compares precision on the same 23 senses that have WordNet examples. It compares recall on all 73 senses. It compares Kappa on the 23 WordNet examples and the sample of generated examples the judges rated.

		Generated	WordNet
Yield		92%	34%
Senses with examples		67	23
Avg. words in context		5.91	7.87
Precision (same 23)	Fine	36%	52%
	Coarse	40%	67%
Recall	Fine	25%	18%
	Coarse	30%	23%
Fleiss' Kappa	Fine	0.43	0.39
	Coarse	0.48	0.49

Table 1: Generated examples vs. WordNet

Errors occur when 1) the corpus is missing a word sense; 2) LDA fails to find good sense indicators; or 3) Context Generation fails to generate a sense-appropriate context.

Our method succeeds when (1) the target sense occurs in the corpus, (2) LDA finds good indicators for it, and (3) Context Generation uses them to construct a sense-appropriate context. For example, the first sense of *advantage* is "the quality of having a superior or more favorable position," for which we obtain the sense indicators *support, work, time, order, life, knowledge, mind, media, human, market, experience, nature, make, social, information, child, individual, cost, people, power, good, land, strategy,* and *company*, and generate (among others) the context *...**knowledge** gave him an advantage….*

Errors occur when any of these 3 steps fails. Step 1 fails for the sense "reduce in scope while retaining essential elements" of *reduce* because it is so general that no good example exists in the corpus for it. Step 2 fails for the sense of *force* in "the force of his eloquence easily persuaded them" because its sense indicators are *men, made, great, page, man, time, general, day, found, side, called,* and *house*. None of these words are precise enough to convey the sense. Step 3 fails for the sense of *advantage* as "(tennis) first point scored after deuce," with sense indicators *point, game, player, tennis, set, score, points, ball, court, service, serve, called, win, side, players, play, team, games, match, wins, won, net, deuce, line, opponent,* and *turn*. This list looks suitably tennis-related. However, the generated context *…the **player** has an advantage…* fits the first sense of *advantage*; here the indicator *player* for the tennis sense is misleading.

4 Contributions and Limitations

This paper presents what we believe is the first system for generating sense-appropriate contexts to illustrate different word senses even if they have the same part of speech. We define the problem of generating sense-targeted contexts for vocabulary learning, factor it into Sense Indicator Extraction and Context Generation, and compare the resulting contexts to WordNet in yield, precision, and recall according to human judges who decided, given definitions of all senses, which one(s) fit each context, without knowing its source or intended sense. This test is much more stringent than just deciding whether a given word sense fits a given context.

There are other possible baselines to compare against, such as Google snippets. However, Google snippets fare poorly on criteria for teaching children vocabulary (Liu *et al.* under revision). Another shortcoming of this alternative is the inefficiency of retrieving all contexts containing the target word and filtering out the unsuitable ones. Instead, we compile constraints on suitability into a generator that constructs only contexts that satisfy them. Moreover, in contrast to retrieve-and-filter, our constructive method (concatenation of overlapping Google 5-grams) can generate novel contexts.

There is ample room for future improvement. We specify word senses as WordNet synsets rather than as coarser-grain dictionary word senses more natural for educators. Our methods for target word document corpus construction, Sense Indicator Extraction, and Context Generation are all fallible. On average, 1 of 3 human judges rated the resulting contexts as sense-appropriate, half as many as for WordNet examples. However, thanks to high yield, their recall surpassed the percentage of synsets with WordNet examples. The ultimate criterion for evaluating them will be their value in tutorial interventions to help students learn vocabulary.

Acknowledgments

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080157 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We thank the reviewers and our judges.

References

Isabel L. Beck, Margaret G. Mckeown and Linda Kucan. 2002. Bringing Words to Life: Robust Vocabulary Instruction. NY, Guilford.

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

Weisi Duan and Alexander Yates. 2010. Extracting Glosses to Disambiguate Word Senses. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles.

Ruben G. Fukkink, Henk Blok and Kees De Glopper. 2001. Deriving word meaning from written context: A multicomponential skill. Language Learning **51**(3): 477-496.

Patricia A. Herman, Richard C. Anderson, P. David Pearson and William E. Nagy. 1987. Incidental acquisition of word meaning from expositions with varied text features. Reading Research Quarterly **22**(3): 263-284.

Joseph R. Jenkins, Marcy Stein and Katherine Wysocki. 1984. Learning vocabulary through reading. American Educational Research Journal **21**: 767-787.

Melanie R. Kuhn and Steven A. Stahl. 1998. Teaching children to learn word meaning from context: A synthesis and some questions. Journal of Literacy Research **30**(1): 119-138.

Anagha Kulkarni and Ted Pedersen. 2005. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. Proceedings of the Second Indian International Conference on Artificial Intelligence, Pune, India.

Liu Liu, Jack Mostow and Greg Aist. 2009. Automated Generation of Example Contexts for Helping Children Learn Vocabulary. Second ISCA Workshop on Speech and Language Technology in Education (SLaTE), Wroxall Abbey Estate, Warwickshire, England.

Liu Liu, Jack Mostow and Gregory S. Aist. under revision. Generating Example Contexts to Help Children Learn Word Meaning. Journal of Natural Language Engineering.

David Martinez, Oier Lopez de Lacalle and Eneko Agirre. 2008. On the use of automatically acquired examples for all-nouns word sense disambiguation. Journal of Artificial Intelligence Research **33**: 79--107.

Rada Mihalcea. 2002. Bootstrapping large sense tagged corpora. Proceedings of the 3rd International Conference on Languages Resources and Evaluations LREC 2002, Las Palmas, Spain.

R. Uslan Mitkov, Le An Ha and Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple choice test items. Natural Language Engineering **12**(2): 177-194.

William E. Nagy, Richard C. Anderson and Patricia A. Herman. 1987. Learning Word Meanings from Context during Normal Reading. American Educational Research Journal **24**(2): 237-270.

William E. Nagy, Patricia A. Herman and Richard C. Anderson. 1985. Learning words from context. Reading Research Quarterly **20**(2): 233-253.

Roberto Navigli, Kenneth C. Litkowski and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained English all-words task. Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics**:** 30-35.

Juan Pino and Maxine Eskenazi. 2009. An Application of Latent Semantic Analysis to Word Sense Discrimination for Words with Related and Unrelated Meanings. The 4th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2009 Workshops, Boulder, CO, USA.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Elinore K. Schatz and R. Scott Baldwin. 1986. Context clues are unreliable predictors of word meanings. Reading Research Quarterly **21**: 439-453.

Paula J. Schwanenflugel, Steven A. Stahl and Elisabeth L. Mcfalls. 1997. Partial Word Knowledge and Vocabulary Growth during Reading Comprehension. Journal of Literacy Research **29**(4): 531-553.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin **86**(2): 420-428.

Mark Steyvers and Tom Griffiths. 2006. Probabilistic topic models. Latent Semantic Analysis: A Road to Meaning. T. Landauer, D. McNamara, S. Dennis and W. Kintsch. Hillsdale, NJ, Laurence Erlbaum.

David Yarowsky. 1995. Unsupervised WSD rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Massachusetts Institute of Technology, Cambridge, MA.