Coarse-grained +/-Effect Word Sense Disambiguation for Implicit Sentiment Analysis

Yoonjung Choi, Janyce Wiebe, and Rada Mihalcea

Abstract—Recent work has addressed opinion inferences that arise when opinions are expressed toward +/-effect events, events that positively or negatively affect entities. Many words have mixtures of senses with different +/-effect labels, and therefore word sense disambiguation is needed to exploit +/-effect information for sentiment analysis. This paper presents a knowledge-based +/-effect coarse-grained sense disambiguation method based on selectional preferences modeled via topic models. The method achieves an overall accuracy of 0.83, which represents a significant improvement over three competitive baselines.

Index Terms—Sentiment Analysis, Implicit Sentiment, Word Sense Disambiguation, Opinion Inference.

1 Introduction

Sentiment analysis extracts opinions from many kinds of texts such as reviews, news, and social media messages and has been exploited in many applications such as review mining, election analysis, and question answering. Research in sentiment analysis has plateaued at a somewhat superficial level, providing methods that exhibit a fairly shallow understanding of subjective language as a whole. In particular, past research in natural language processing has mainly addressed explicit opinion expressions [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], ignoring implicit opinions expressed via implicatures, i.e., default inferences.

Recently, to determine implicit opinions, Deng et al. [16], [17], [18] address a type of opinion inference that arises when opinions are expressed toward events that have positive or negative effects on entities, called *+/-effect* events.[1] Deng et al. show how sentiments toward one entity may be propagated to other entities via *opinion inference rules*, improving the coverage and accuracy of sentiment analysis. They give the following example:

(1) *The bill would curb skyrocketing health care costs.*

The writer expresses an explicit negative sentiment (by *skyrocketing*) toward the entity, *health care costs*. The existing sentiment analysis system can determine it. However, the existing explicit sentiment analysis system cannot determine the sentiment toward *the bill*. With opinion inference rules, not only the sentiment toward *health care costs* but also the sentiment toward *the bill* can be inferred. The event, *curb*, has a negative effect (i.e., -effect) on *skyrocketing health care costs*, since they are reduced. We can reason that the writer is positive toward the event because it has a negative effect on *costs*, toward which the writer is negative. From there, we can reason that the writer is positive toward *the bill*, since it conducts the positive event.

Now, consider another example:

(2) *Oh no! The voters passed the bill.*

Here, the writer expresses an explicit negative sentiment toward the passing event because of the expression *Oh no!*. Although we cannot know the sentiment toward *the bill* with an existing sentiment analysis system, we can infer it with opinion inference rules. The passing event has a positive effect (i.e., +effect) on *the bill* since it brings it into existence. Since the writer is negative toward an event that benefits *the bill*, we can infer that the writer is negative toward *the bill* itself.[2]

The ultimate goal is to develop a fully automatic system capable of recognizing such inferred attitudes. The system will require a set of implicature rules and an inference mechanism. In [17], Deng et al. present a graph-based model in which inference is achieved via propagation. They show that such inferences may be exploited to significantly improve explicit sentiment analysis systems.

To achieve its results, the inference system needs the ability to recognize +/-effect events in text. Deng et al. [18] include such an extraction component, but it simply looks for the presence of words in a lexicon. Such an approach is not effective because, as Choi et al. [19], [20] discovered, there is substantial sense ambiguity – words often have mixtures of +effect, -effect, and Null (i.e., neither) senses. In fact, they found that 45.6% verbs in WordNet contain two or more senses (i.e., homonymy). Among them, 63.8% words have some kind of +/-effect ambiguity (11.3% words have mixtures of +effect, -effect, and Null senses; 3.9% words have mixtures of +effect and -effect; 25.9% and 22.7% words have +effect & Null and -effect & Null). Thus, Choi et al. [19], [20] built a sense-level +/-effect lexicon, called EffectWordNet.

The sense of the word in context affects whether (or which) inference should be made. The meaning of *pass* in (2) is the following:

Manuscript received February, 2017; revised June, 2017.

1. Some of those papers use the terms *goodFor/badFor*.

2. As addressed in [18], such inferences are defeasible. They capture this in their ILP model by including slack variables that allow the rules to be violated.

S_1: (v) legislate, pass (make laws, bills, etc. or bring into effect by legislation) **+effect**

Under this sense, *pass* is, in fact, +effect for its theme. But consider (3):

(3) *Oh no! They passed the bridge.*

The sense of *pass* in this context is:

S_2: (v) travel by, pass by, surpass, go past, go by, pass (move past) **Null**

This type of passing event does not (in itself) positively or negatively affect the thing passed (*bridge*). This use of *pass* does not warrant the inference that the writer is negative toward the bridge.

The following is another example of a word with senses of different classes:

carry:
S_3: (v) carry (keep up with financial support) "The Federal Government carried the province for many years" **+effect**
S_4: (v) carry (capture after a fight) "The troops carried the town after a brief fight" **-effect**

In the first sense S_3, *carry* has positive polarity toward *the province*. However, in the second sense S_4, *carry* has negative polarity toward *the town*, since it is captured by the troops. Moreover, although a word may not have both +effect and -effect senses, it may have mixtures of (+effect or -effect) and Null similar to *pass*. These examples illustrate that exploiting +/-effect event information for sentiment inference requires word sense disambiguation (WSD).

Therefore, our paper focuses on **+/-effect WSD**, which is important for opinion inferences to extract implicit opinions. The goal of this paper is to show that we can effectively identify the +/-effect events in a given text. Since our task is new, the architecture is different from typical WSD systems.

In this paper, we address the following task: given +/-effect labels of *senses*, determine whether an instance of a word in the corpus is being used with a +effect, -effect, or Null sense. Consider a word *W* that contains seven senses, where senses $\{S_1, S_3, S_7\}$ are -effect; $\{S_2\}$ is +effect; and $\{S_4, S_5, S_6\}$ are Null. For our purposes, we do not need to perform fine-grained WSD to pinpoint the exact sense; to recognize that an instance of W is -effect, for example, the system only needs to recognize that W is being used with *one* of the senses $\{S_1, S_3, S_7\}$. Thus, we can perform coarse-grained WSD, which is often more tractable than fine-grained WSD.

Though supervised WSD is generally the most accurate method, we do not pursue a supervised approach, because the amount of available sense-tagged data is limited. Instead, we conduct a knowledge-based WSD method that exploits WordNet[3] relations and glosses. We use sense-tagged data (SENSEVAL-3) only as gold-standard data for evaluation.

3. WordNet 3.0, http://wordnet.princeton.edu/

Our WSD method is based on *selectional preferences*, which are preferences of verbs to co-occur with certain types of arguments [21], [22], [23]. We hypothesized that preferences would be fruitful for our task, because +/-effect is a semantic property that involves affected entities. Consider the following WordNet information for *climb*:

climb:
S_1: (v) climb, climb up, mount, go up (go upward with gradual or continuous progress) "Did you ever climb up the hill behind your house?" **Null**
S_2: (v) wax, mount, climb, rise (go up or advance) "Sales were climbing after prices were lowered" **+effect**
S_3: (v) climb (slope upward) "The path climbed all the way to the top of the hill" **Null**
S_4: (v) rise, go up, climb (increase in value or to a higher point) "prices climbed steeply"; "the value of our house rose sharply last year" **+effect**

Senses S_1 & S_3 are both Null. We expect them to co-occur with *hill* and similar words such as *ridge* and *mountain*. And, we expect such words to be more likely to co-occur with S_1 & S_3 than with S_2 & S_4. Senses S_2 & S_4 are both +effect, since the affected entities are increased. We expect them to co-occur with *sales*, *prices*, and words similar to them. And, we expect such words to be more likely to co-occur with S_2 & S_4 than with S_1 & S_3. This example illustrates the motivation for using selectional preferences for +/-effect WSD.

We model sense-level selectional preferences using topic models, specifically Latent Dirichlet Allocation (LDA) [24]. We utilize LDA for modeling relations between sense groups and their arguments, and then carry out coarse-grained +/-effect WSD by comparing the topic distributions of a word instance and candidate sense groups and choosing the sense group that has the highest similarity value. Because selectional preferences are preferences toward arguments, the method must create a set of arguments to consider for each sense group. We exploit information in WordNet for automatically defining sets of arguments.

The system carries out WSD by matching word instances to sense groups. While the obvious way to group senses is simply by +/-effect label, the system does not need to group senses in this way. We experiment with a clustering process that allows more than one sense group with the same label for a given word. The motivation for allowing this is that there may be subsets of senses that have the same +/-effect label, but which are more similar to each other than they are to the other senses with the same +/-effect label. We also experiment with using mixtures of manually and automatically assigned sense labels in this clustering process, exploiting the results presented in [19] for automatically assigning +/-effect labels to verb senses in WordNet.

In the remainder of this paper, related work is first discussed in Section 2 and the task is defined in Section 3. The method for creating the WSD system is described in Section 4. The experiments and results are presented in Section 5. Section 6 presents the conclusions.

2 Related Work

+/-Effect events were first described in [16]. They identify several varieties of +/-effect events, including creation/destruction (changes in states involving existence), gain/loss (changes in states involving possession), and benefit/injury.[4]

In the general area of sentiment analysis, researchers have developed various lexicons involving polarities: sentiment lexicons (e.g., [25]), connotation lexicons (e.g., [26]), and +/-effect lexicons (e.g., [19]). Interestingly, though these polarities are related, they are not the same. Choi et al. [20] give the following example:

> **perpetrate**:
> S: (v) perpetrate, commit, pull (perform an act, usually with a negative connotation) "perpetrate a crime"; "pull a bank robbery"

This sense of *perpetrate* has a negative connotation, and is an objective (i.e., non-sentiment-bearing) term in `SentiWordNet`. However, it has a positive effect on the object, *a crime*, since performing a crime brings it into existence. Like this, a single event may have different polarities of sentiment, connotation, and +/-effect. Therefore, we need to acquire a new type of lexicon of +/-effect events to make opinion inference.

While there has been much work developing sentiment and related lexicons, much less work has addressed resolving the relevant sense ambiguities in context, as we do in this paper. Akkaya et al. [27], [28], [29] perform coarse-grained WSD for sentiment-bearing versus non-sentiment bearing word instances. However, their method is supervised per-word WSD, and thus requires coarse-grained sense-tagged training data for each word. Xia et al. [30] adopt a Bayesian model to resolve the polarity ambiguity.

Turning to WSD in general, supervised WSD approaches (e.g., [31], [32], [33]) generally show the best accuracy, but require sense-tagged training data. Unsupervised (e.g., [34], [35], [36]) and knowledge-based approaches that rely on the use of external lexical resources such as dictionaries and ontologies (e.g., [37], [38], [39], [40]) typically have better coverage, since they do not require sense-tagged data.

The WSD research most relevant to ours involves selectional preferences and topic models. [41], [42] utilize selectional preferences for WSD by obtaining them as sets of disjoint classes across WordNet hypernyms. They demonstrate a small, but positive improvement on WSD. We hypothesized that preferences would be even more beneficial for our WSD task, since the +/-effect property is defined in terms of affected entities. Topic models have been used for WSD [43], [44], [45], but not to model selectional preferences for WSD as we do. Note that LDA has been used for selectional preferences [46], [47], but only to handle word-level predicates, so the methods are not directly applicable for WSD. Our method is a novel use of LDA to model selectional preferences for WSD.

4. Their annotation manual and data are available at http://mpqa.cs.pitt.edu/.

3 Task Definition

The task addressed in this paper is to recognize whether word instances in a corpus are used with +effect, -effect, or Null senses. Specifically, the gold standard consists of pairs $\langle w, l \rangle$, where w is an instance of word W in the corpus, and l is w's label, meaning that w is a use of W with a sense whose label is l. In this work, the gold standard is created by combining sense-tagged (SENSEVAL-3) data and +/-effect sense labels as follows: $\langle w, l \rangle$ in our gold standard means that w has sense label W_s, and W_s has +/-effect label l.

For example, the label for the instance of *pass* in (2) is +effect, because the sense is S_1, and S_1 has the label +effect.

4 Word Sense Disambiguation

This section describes our method for building a selectional-preference +/-effect coarse-grained WSD system, given a resource such as WordNet and +/-effect labels on word senses.

In the first step, a coarse-grained sense inventory is constructed, by grouping senses (Section 4.1). The ultimate WSD system will assign each word instance in the corpus to one of the sense groups. For final evaluation, a word instance w that the WSD system has assigned to any sense group with label l is mapped to the pair $\langle w, l \rangle$, for comparison with the gold standard. The obvious grouping is simply by +/-effect labels: one group for the +effect senses, one for the -effect senses, and one for the Null senses. Alternatively, there may be multiple groups for a single label, where the senses in a group are more closely related to each other than they are to other senses with the same label. Our hypothesis for experimenting with variable grouping (i.e., allow more than one sense group with the same label for a given word) is that an effective method could be developed for creating a more fine-grained sense inventory customized to our task that would result in more accurate WSD performance.

Once the sense inventory is created, a model of selectional preferences for the sense groups is developed. Selectional preferences are preferences toward arguments. Thus, we have to identify a set of arguments for each group (Section 4.2). For example, suppose that S_2 and S_4 of *climb* are one sense group. The arguments for this group include nouns extracted from their glosses (*sales*, *prices*, etc.) together with others found by WordNet relation expansion. The final step in creating the WSD system is to model relations between sense groups and arguments to capture selectional preferences using LDA modeling (Section 4.3). This step defines argument class distributions, where the classes are hidden variables.

Finally, these distributions are exploited to perform WSD, as described in Section 4.4.

4.1 Sense Grouping

Performing coarse-grained WSD has the advantage that individual senses are aggregated, providing more information about each coarse-level sense.

For each word, senses can be simply grouped by label. However, a problem is that senses with the same +/-effect label but with very different selectional preferences are forced into the same group, making them indistiguishable to the WSD system. For instance, one sense of *carry* is *win in an election* and another is *keep up with financial support*. Though both are +effect, they have very different arguments. Nevertheless, they are forced into the same group. Because such groups contain several types of arguments, they can confuse the LDA models.

Thus, we adopt sense clustering and allow multiple groups with the same label, which can benefit the LDA models because each sense group can have purer arguments. The process is as follows: first features are extracted from WordNet, then senses are clustered based on the features, and finally labels are assigned to clusters.

The features represent the absence or presence of the following words: words in the synset and the gloss for sense S_i; words in the synsets and the glosses for all S_i's hypernyms (i.e., more general word senses); words in the synsets and the glosses of S_i's troponyms (i.e., more specific word senses); words in the synsets and the glosses of S_i's verb groups (i.e., verb senses with similar meanings).

For sense clustering, we adopt expectation maximization (EM) [48] as implemented in the Weka library,[5] which is modeled as a mixture of Gaussians. It follows an iterative approach to find the parameters of the probability distribution. In each iteration, the E-step (Expectation) estimates the probabilities of each data belong to each cluster, and the M-step (Maximization) estimates the parameter of the probability distribution of each cluster. In Weka, EM assigns a probability distribution to each instance, the probability of it belonging to each cluster. Further, EM selects the number of clusters automatically by maximizing the log-likelihood. It begins with one cluster and continues to add clusters until the estimated log-likelihood is decreased.

After clustering, labels are assigned to clusters as follows. If all or a majority of senses in a cluster have the same label, then the cluster is assigned that label. If there is not a majority, then the cluster is labeled Null. That is, if a cluster consists of three +effect senses and one Null sense, the cluster is assigned the +effect. However, if a cluster contains two +effect senses and two -effect senses, this cluster is labeled Null. In this case, some Null labels are assigned to clusters where there are no Null elements inside. However, since our task is to identify words that are likely to be +/-effect, we want to focus on +effect and -effect labels rather than the Null class.

4.2 Arguments for Selectional Preferences

After grouping senses, arguments for each sense group must be extracted to exploit selectional preferences. Gloss information (definitions and examples) and semantic relations in WordNet are utilized.

We first combine gloss information of all senses in each sense group SG_k. Since glosses are not long, we consider all nouns in the combined glosses as arguments of the given sense group. We call this noun set N.

5. Weka3, http://www.cs.waikato.ac.nz/ml/weka/

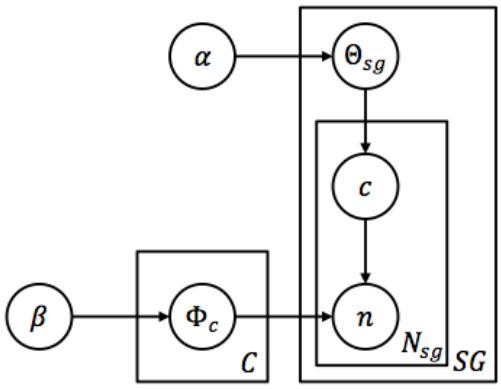

Fig. 1. Plate notation representing our topic model.

We also consider arguments gleaned from senses related to those in the sense group. While such arguments are less tightly coupled to the senses they are being extracted for, we hypothesized that, on balance, having a larger number of arguments may improve overall performance.

Let $commonSynset(word1, word2)$ be $True$ if there is at least one synset that contains a sense of $word1$ and a sense of $word2$. We add all words new for which $commonSynset(n, new)$ for some $n \in N$. The synset relation is the closest relationship between senses in WordNet, so we anticipated that adding these new arguments would be a conservative way to increase recall.

Going one step further, we consider WordNet verb relations for sense S_i in each sense group SG_k because we hypothesize that the super-subordinate relations can provide richer information. All nouns in glosses of hypernyms and troponyms of S_i are extracted and added to the argument set. In addition, the argument set contains all nouns in glosses of the senses that are in the same verb group with S_i. Generally speaking, the coverage of WordNet verb groups is not large, but the relations are reliable.

4.3 Topic Model

To model relations between sense groups and arguments for each +/-effect event, we adopt LDA, which is a generative model that discovers similarities in data using latent variables. It was introduced to model a set of documents in terms of topics, representing the underlying semantic structure of a document collection. In this paper, sense groups play the role of documents, arguments play the role of terms, and argument classes play the role of topics in traditional usage of LDA. That is, rather than modeling relations between documents and terms, we model relations between sense groups and arguments. One advantage of LDA is argument classes need not be pre-defined, since LDA discovers these classes automatically. We adopt a variant of LDA suggested by Griffiths and Steyvers [49], [50], [51].

Figure 1 shows the graphical model of our proposed topic model. Arrows represent conditional dependencies between variables. SG is a set of sense groups, N_{sg} is a set of arguments for each sense group sg, and C is a set of argument classes, which are hidden variables being discovered by the model.

Each sense group sg has a corresponding multinomial distribution Θ_{sg} over latent argument classes c. Distribution Θ_{sg} is defined from a Dirichlet distribution with prior parameter α. Each argument class c also has a corresponding multinomial distribution Φ_c over arguments n. Distribution Φ_c is defined from a Dirichlet with prior parameter β. To generate an argument n, a hidden argument class c is first chosen by Θ_{sg}, and then an argument n is chosen from Φ_c. The formal process is as follow:

1) Choose $\Theta_{sg} \sim Dir(\alpha)$, where $sg \in SG$ and $Dir(\alpha)$ is the Dirichlet distribution for parameter α.
2) Choose $\Phi_c \sim Dir(\beta)$, where $c \in C$.
3) To generate an argument,
   a) Draw a specific argument class $c \sim \Theta_{sg}$
   b) Draw an argument $n \sim \Phi_c$

In this model, the main variables are the argument distribution Φ for each argument class and the argument class distribution Θ for each sense group. They can be estimated directly, but this approach can get stuck in a local maximum of the likelihood function. Another method is to directly estimate the posterior distribution over argument class c [52]. For posterior inference, we use Gibbs sampling, which has been shown to be a successful inference method for LDA [51]. It sequentially samples variables from their distribution conditioned on the current values of all other variables. With these samples, we can approximate the posterior distribution. For the implementation, we use the Mallet library, and use its default setting that assumes seven topics.[6]

4.4 Word Sense Disambiguation

The topic model defines argument class distributions for each sense group. Let D_k be the argument class distribution of SG_k.

To disambiguate word instance W in the corpus, the nouns within a window size of five are extracted to serve as its arguments. [7] We create a test instance with these nouns and obtain the argument class distribution of W by the topic model described above. Let this distribution be D_W.

We hypothesized that arguments can help determine the +/-effect polarity of senses for the given word. Each word can have several meanings, and the polarity can be different according to the meanings. We can distinguish these meanings based on their arguments. That is, our assumption is that if senses of W have similar types of arguments, they have the same +/-effect polarity. Thus, the system chooses the sense group that has the highest similarity value to D_W, since similar types of arguments can be expected to show similar argument class distributions. In particular, similarity is assessed as the cosine value between the distribution vectors D_W and D_k, for all D_k, and the k for which similarity is highest is selected. That is, D_W has higher similarity value to D_3 than the others (D_1 and D_2), then SG_3 is selected. Finally, W is assigned the label of SG_k as its +/-effect label.

5 Experiments

In this section, we first describe our data (Section 5.1) and baselines for comparison (Section 5.2). We provide our experimental results (Section 5.3), and then we present the role of word sense clustering (Section 5.4). Finally, we show the role of manual (vs. automatic) +/-effect sense labels (Section 5.5).

5.1 Gold Standard Data and Sense Annotations

For evaluation, the SENSEVAL-3 English lexical sample task data is used.[8] It provides training and test data for 57 words out of which 32 are verbs. Since we consider only verbs as +/-effect events in this paper, we only utilize the verb data. We adopt the SENSEVAL-3 test data as our test data, which has a total of 1,978 instances for the 32 verbs.

To complete the gold standard, +/-effect labels are required. Although Choi et al. [19] provide their annotated data, that data does not include the 32 verbs in the Senseval-3 data. Thus, we manually annotate the senses of all 32 verbs as +effect, -effect, or Null. The total number of senses is 246. We followed the annotation scheme in [16], [19], which was found to lead to good inter-annotator agreement (0.84 percent agreement and 0.75 κ value reported in their study). Our annotation rate was approximately 100 senses per hour. Note that sense labeling requires much less effort than creating sense-tagged training data, and can be viewed as a manual augmentation of WordNet, which was itself manually created. For future additional annotations, [19] give a method for guided manual annotation, where the model identifies unlabeled words that are likely to have +/-effect senses.

According to the manual annotations, among 246 senses, 49 senses (19.9%) are +effect, 36 (14.6%) are -effect, and the rest are Null. Among 32 verbs, 2 have +effect, -effect, and Null senses, and 20 have Null and one of +/-effect senses. Thus, we see that 68.75% (22/32) of the verbs chosen for inclusion in SENSEVAL-3 require sense disambiguation to determine +/-effect labels for word instances.

Based on the sense labels, labels are assigned to the SENSEVAL-3 data to create the gold standard used as test data in all the experiments reported in this paper. The test data consists of 467 +effect, 108 -effect, and 825 Null instances. Importantly, note that since our proposed method is a knowledge-based WSD approach, we do not need any training data.

5.2 Baselines

As one baseline system, we adopt WordNet::SenseRelate::TargetWord,[9] which is an unsupervised WSD method that is freely available [34]. In the table

6. Mallet, http://mallet.cs.umass.edu/topics.php
7. The window size is simply a heuristic.
8. SENSEVAL-3, http://www.senseval.org/
9. WordNet::SenseRelate, http://senserelate.sourceforge.net/

TABLE 1
Experimental results for *All* (all 32 verbs) and *Conf* (22 verbs with +/-effect ambiguity) set.

		Majority		BL1:SenseRelate		BL2:GWSD		Our Method	
		All	**Conf**	**All**	**Conf**	**All**	**Conf**	**All**	**Conf**
Accuracy		0.701	0.625	0.535	0.519	0.499	0.425	**0.880**	**0.833**
+effect	Precision	0.000	0.000	**0.807**	**0.814**	0.568	0.534	0.791	0.776
	Recall			0.449	0.469	0.368	0.344	**0.808**	**0.794**
	F-measure			0.577	0.595	0.447	0.418	**0.799**	**0.785**
-effect	Precision	0.000	0.000	0.620	0.438	0.556	0.410	**0.943**	**0.921**
	Recall			0.220	0.130	0.425	0.313	**0.817**	**0.759**
	F-measure			0.325	0.200	0.482	0.355	**0.875**	**0.832**
Null	Precision	0.701	0.625	0.804	0.773	0.834	0.736	**0.909**	**0.856**
	Recall	**1.000**	**1.000**	0.606	0.650	0.550	0.477	0.914	0.864
	F-measure	0.824	0.769	0.691	0.706	0.663	0.579	**0.911**	**0.860**

of results, this system is referred to as *BL1:SenseRelate*. Because it performs unsupervised WSD, it does not require sense-tagged training data. Since it is a WSD method, its output is a sense. Thus, after running it, we assign +/-effect labels based on the manually annotated senses described above. Among 1,978 instances in the test data, it does not provide any sense information for 691 instances (34.93%).

Another system is GWSD (*BL2:GWSD*), which is an unsupervised graph-based WSD system developed by Sinha and Mihalcea [53].[10] Since its output is also a sense, we assign +/-effect labels based on the manually annotated senses similar to the strategy used for the previous baseline. When we run GWSD, we select the verbs as the target part of speech, Leacock & Chodorow (*lch*) as the similarity metric used to build the graph, six for the window size, and *indegree* for the graph centrality measure (indegree was found to have a performance comparable to other more sophisticated measures, and it is more efficient).

The other baseline system, called *Majority* Baseline simply chooses the majority class (Null).

5.3 Experimental Results

We evaluate our system for two verb sets: *All* consists of all 32 verbs and *Conf* contains the 22 verbs with +/-effect ambiguity.

Table 1 shows results for the *Majority* baseline, *BL1:SenseRelate*, *BL2:GWSD*, and our system. It gives accuracy, precision (P), recall (R), and f-measure (F) for all three labels.

While *BL1:SenseRelate* has the highest +effect precision and *Majority* baseline has the highest Null recall (as it assigns everything to the Null class), our system is substantially better on all other measures.

As we mentioned in Section 5.2, two baseline systems (except *Majority*) did not detect any sense information for many instances, so their recall is low. Nevertheless, they show high +effect and Null precision. In addition, in *BL2:GWSD*, the accuracy is quite good.

10. GWSD, https://web.eecs.umich.edu/~mihalcea/downloads.html

However, our system outperforms them. It shows high recall scores for all three labels and the best accuracy score. Moreover, our system is better at detecting -effect events than all three baselines. In fact, the overall accuracy is 0.83 and all three f-measures are over 0.78, representing a good performance for a WSD approach that is not supervised.

Table 2 and Table 3 shows the role of argument types. As we explained in Section 4.2, we utilize gloss information and semantic relations in WordNet to extract arguments for selectional preferences. All cases of arguments are as follows:

- **ArgSet1**: All nouns (*Ns*) in gloss information of senses S in each sense group.
- **ArgSet2**: Synsets of *Ns*.
- **ArgSet3**: All nouns in gloss information of hypernyms of S.
- **ArgSet4**: All nouns in gloss information of troponyms of S.
- **ArgSet5**: All nouns in gloss information of verb groups of S.

Table 2 presents the performance of each argument type and all of them. Based on our experiments, we get the best result with the combination of ArgSet1, ArgSet2, and ArgSet5. Table 3 shows the results of backward-ablation. We can know that each argument type is helpful to our task even though the difference is not big.

5.4 The Role of Word Sense Clustering

As described above, sense groups can be simply grouped by label. That is, each word has one sense group for each label. In this case, each word can have at most 3 groups: +effect, -effect, and Null. We call this method the fixed sense grouping.

Table 4 shows the result of the fixed sense grouping (Fixed) based on manually annotated senses described in Section 5.1. It also includes results for full fine-grained WSD

TABLE 2
Performance of argument types on the *Conf* (22 verbs with +/-effect ambiguity) set.

	+effect			-effect			Null			Accuracy
	P	R	F	P	R	F	P	R	F	
ArgSet1	0.775	0.794	0.784	0.921	0.759	0.832	0.856	0.863	0.860	0.832
ArgSet2	0.773	0.791	0.782	0.921	0.759	0.832	0.855	0.862	0.858	0.831
ArgSet3	0.767	0.791	0.779	0.921	0.759	0.832	0.854	0.857	0.856	0.828
ArgSet4	0.726	0.804	0.763	0.921	0.759	0.832	0.855	0.822	0.838	0.811
ArgSet5	0.772	0.836	0.803	0.921	0.759	0.832	0.876	0.854	0.865	0.841
ArgAll(ArgSet1-5)	0.776	0.794	0.785	0.921	0.759	0.832	0.856	0.864	0.860	0.833
Best(ArgSet1,2,5)	0.778	0.838	0.807	0.921	0.759	0.832	0.877	0.858	0.868	0.844

TABLE 3
The results of backward-ablation on the *Conf* (22 verbs with +/-effect ambiguity) set.

	+effect			-effect			Null			Accuracy
	P	R	F	P	R	F	P	R	F	
ArgAll(ArgSet1-5)	0.776	0.794	0.785	0.921	0.759	0.832	0.856	0.864	0.860	0.833
ArgAll - ArgSet1	0.766	0.796	0.781	0.921	0.759	0.832	0.856	0.856	0.856	0.829
ArgAll - ArgSet2	0.755	0.800	0.777	0.921	0.759	0.832	0.857	0.847	0.852	0.825
ArgAll - ArgSet3	0.770	0.814	0.791	0.921	0.759	0.832	0.865	0.856	0.861	0.835
ArgAll - ArgSet4	0.773	0.812	0.792	0.921	0.759	0.832	0.864	0.858	0.861	0.836
ArgAll - ArgSet5	0.768	0.810	0.788	0.921	0.759	0.832	0.863	0.855	0.859	0.833

TABLE 4
Comparison among fine-grained WSD (No Groups), a fixed number of sense groups (Fixed), and a variable number of sense groups (Our Method) on *Conf* set.

		No Group	Fixed	Our Method
Accuracy		0.585	0.758	**0.833**
+effect	P	0.502	0.689	0.776
	R	0.699	0.743	0.794
	F	0.584	0.715	**0.785**
-effect	P	0.500	0.638	0.921
	R	0.798	0.815	0.759
	F	0.615	0.716	**0.832**
Null	P	0.713	0.824	0.856
	R	0.490	0.760	0.864
	F	0.581	0.791	**0.860**

(No Group). The same gold standard test set continues to be used for all experiments and only the *Conf* set is evaluated.

As expected, accuracy and all f-measures are the worst for fine-grained WSD, where no sense grouping is performed. Also, accuracy and all f-measures are substantially better than Fixed after automatically refining the system's sense inventory via clustering.

Following is an example illustrating how clustering can improve performance. Consider *suspend*, which has 5 -effect senses and 1 Null sense. Following are examples from SENSEVAL-3. The sense in Ex1-Ex2 is S_3, *bar temporarily*. The sense in Ex3-Ex4 is S_5, *make inoperative or stop*.

(Ex1) S_3: He was later suspended for two European games for unsporting behaviour.
(Ex2) S_3: He was suspended for two years after he tested positive for drugs when finishing second in the 1988 New York race.
(Ex3) S_5: France is to suspend nuclear tests at its South Pacific atoll site, Mururoa, this year, M Pierre Beregovoy, Prime Minister, said in his inaugural speech to parliament yesterday.
(Ex4) S_5: That was good enough to prompt Gordon Taylor, the PFA chief executive, to suspend the threat of industrial action.

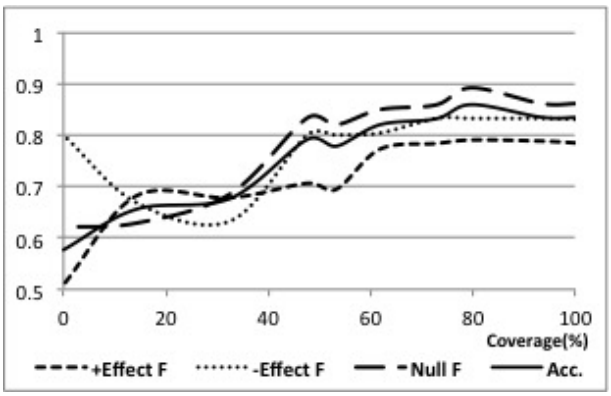

Fig. 2. Learning curve on *Conf* with increasing percentages of manual sense annotations.

S_3 and S_5 are both -effect, so a fixed sense grouping forces them into the same group. But the contexts in which S_3 and S_5 are used are different, and the topic model must contend with one -effect group which includes quite varied contexts (sports related, politics related, etc). In fact, the system incorrectly labels Ex3 as Null when the fixed sense groupings are used. With clustering, the system gets all of Ex1-Ex4 correct. A singleton cluster is correctly created for the Null sense (*suspension in a fluid*). S_3 and S_5 are placed into separate groups with other senses. With these purer sense groups, the topic model is able to better model the selectional preferences and provide more accurate results.

5.5 The Role of Manual +/-Effect Sense Labels

Recall that the WSD system assigns the same label to all the senses in a cluster (the majority label, or Null if there isn't one). In Section 5.3 and Section 5.4, we used manually labeled sense data explained in Section 5.1. While sense labeling requires much less labor than sense tagging corpora, it is still desirable not to require full manual sense tagging. In this section, we also utilize EffectWordNet, which automatically labels all verb senses with +/-effect labels [19].[11]

Figure 2 presents a learning curve with increasing percentages of (randomly selected) manual sense labels to determine cluster labels. We only show results for variable sense grouping because we carried out experiments on *Conf* set using 100% automatic labels comparing fixed versus variable sense grouping, and found that performance is much better with variable sense grouping.

On the left, 100% of the labels are automatic. Accuracy is 57.7% which is lower than the 84.4% accuracy reported in Table 1, when 100% of the manual labels are used. The f-measures are lower as well (51 <78.5 for +effect; 80 <83.2 for -effect; and 62 <86.0 for Null). Fortunately, with only 65% of manually annotated senses, we are close to maximum performance; with 80%, we reach the maximum performance. This suggests that, until all verbs have been manually labeled, good performance can still be obtained using some automatic labels to fill out coverage.

11. EFFECTWORDNET, http://mpqa.cs.pitt.edu/lexicons/

6 CONCLUSIONS

In this paper, we investigated a knowledge-based coarse-grained +/-effect WSD approach, which identifies the +/-effect of a word sense based on its surrounding context. Our goal was to show that we can effectively identify the +/-effect events in a given text, which is different from typical WSD systems.

Since our purpose is to determine whether an instance of a word in the corpus is being used with a +effect, -effect, or Null sense, we do not need to perform fine-grained WSD to pinpoint the exact sense. Thus, we performed coarse-grained WSD, which is often more tractable than fine-grained WSD. Moreover, because the amount of available sense-tagged data is limited, we conducted a knowledge-based WSD method, that exploits WordNet relations and glosses, rather than supervised WSD. That is, our method does not require any sense-tagged training data.

The method we proposed relies on selectional preferences. Selectional preferences are modeled using LDA. We used automatic clustering based on the preference arguments, which is extracted from WordNet information, to create a sense inventory customized to our task.

Through several experiments on a test dataset consisting of sense tagged data drawn from SENSEVAL-3, we showed that our method achieves very good performance, with an overall accuracy of 0.83, which represents a significant improvement over three competitive baselines. For the +effect label, even though the precision in one baseline (*BL1:SenseRelate*) is higher than our method, we show a significant improvement in the recall. For the -effect label, our method outperforms all the baseline systems for all the measures. With a majority baseline, since the majority is Null, the recall of Null is 1.0. Although our system has lower recall, we show better precision and f-measure. Also, we show that each argument type used for selectional preferences is helpful to our task.

Moreover, we explored the role of word sense clustering. In our experiments, the variable sense grouping (i.e., allow more than one sense group with the same label) outperforms the fixed sense grouping (i.e., one sense group for each label) and fine-grained WSD (i.e., no grouping). Since it can have purer sense groups with the variable sense grouping, the topic model is able to better model the selectional preferences and provide more accurate results.

Finally, we showed that good performance can still be obtained when automatic labels are used to maximize coverage.

ACKNOWLEDGMENTS

This material is based in part upon work supported by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86. [Online]. Available: http://dx.doi.org/10.3115/1118693.1118704

[2] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 417–424. [Online]. Available: http://dx.doi.org/10.3115/1073083.1073153

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 168–177. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014073

[4] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th International Conference on Computational Linguistics*, ser. COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1220355.1220555

[5] T. Wilson, J. Wiebe, , and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of HLT-EMNLP*, 2005, pp. 347–354.

[6] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 171–180. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242596

[7] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 241–249. [Online]. Available: http://dl.acm.org/citation.cfm?id=1944566.1944594

[8] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 36–44. [Online]. Available: http://dl.acm.org/citation.cfm?id=1944566.1944571

[9] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, October 2013, pp. 1631–1642. [Online]. Available: http://www.aclweb.org/anthology/D13-1170

[10] C. N. dos Santos and M. A. de C. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *COLING*, 2014.

[11] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in *SemEval@COLING*, 2014.

[12] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, 2016.

[13] E. Cambria, N. Howard, Y. Xia, and T. Chua, "Computational intelligence for big social data analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, 2016.

[14] E. Cambria, S. Poria, R. Bajpai, and B. W. Schuller, "Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, 2016, pp. 2666–2677. [Online]. Available: http://aclweb.org/anthology/C/C16/C16-1251.pdf

[15] S. Poria, I. Chaturvedi, E. Cambria, and F. Bisio, "Sentic LDA: improving on LDA with semantic similarity for aspect-based sentiment analysis," in *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*, 2016, pp. 4465–4473. [Online]. Available: https://doi.org/10.1109/IJCNN.2016.7727784

[16] L. Deng, Y. Choi, and J. Wiebe, "Benefactive/malefactive event and writer attitude annotation," in *Proceedings of 51st ACL*, 2013, pp. 120–125.

[17] L. Deng and J. Wiebe, "Sentiment propagation via implicature constraints," in *Proceedings of EACL*, 2014a, pp. 377–385.

[18] L. Deng, J. Wiebe, and Y. Choi, "Joint inference and disambiguation of implicit sentiments via implicature constraints," in *Proceedings of COLING 2014*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, August 2014b, pp. 79–88. [Online]. Available: http://www.aclweb.org/anthology/C14-1009

[19] Y. Choi and J. Wiebe, "+/-effectwordnet: Sense-level lexicon acquisition for opinion inference," in *Proceedings of EMNLP 2014*, 2014a, pp. 1181–1191.

[20] Y. Choi, L. Deng, and J. Wiebe, "Lexical acquisition for opinion inference: A sense-level lexicon of benefactive and malefactive events," in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, 2014b, pp. 107–112.

[21] P. Resnik, "Selectional constraints: an information-theoretic model and its computational realization," *Cognition*, vol. 61, pp. 127–159, 1996.

[22] M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil, "Inducing a semantically annotated lexicon via em-based clustering," in *Proceedings of the 37th ACL*, 1999, pp. 104–111.

[23] T. Van de Cruys, "A neural network approach to selectional preference acquisition," in *Proceedings of EMNLP 2014*, 2014, pp. 26–35.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[25] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of 5th LREC*, 2006, pp. 417–422.

[26] J. S. Kang, S. Feng, L. Akoglu, and Y. Choi, "Connotationwordnet: Learning connotation over the word+sense network," in *Proceedings of the 52nd ACL*, 2014, p. 15441554.

[27] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," in *Proceedings of EMNLP 2009*, 2009, pp. 190–199.

[28] C. Akkaya, J. Wiebe, A. Conrad, and R. Mihalcea, "Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis," in *Proceedings of CoNLL 2011*, 2011, pp. 87–96.

[29] C. Akkaya, J. Wiebe, and R. Mihalcea, "Iterative constrained clustering for subjectivity word sense disambiguation," in *Proceedings of the 14th EACL*, 2014, p. 269278.

[30] Y. Xia, E. Cambria, A. Hussain, and H. Zhao, "Word polarity disambiguation using bayesian model and opinion-level features," *Cognitive Computation*, vol. 7, no. 3, pp. 369–380, 2015. [Online]. Available: https://doi.org/10.1007/s12559-014-9298-4

[31] H. T. Ng, "Getting serious about word sense disambiguation," in *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, 1997, pp. 1–7.

[32] E. Agirre and D. Martínez, "Exploring automatic word sense disambiguation with decision lists and the web," in *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, 2000, pp. 11–19.

[33] G. Tsatsaronis, M. Vazirgiannis, and I. Androutsopoulos, "Word sense disambiguation with spreading activation networks generated from thesauri," in *Proceedings of the 20th IJCAI*, 2007, pp. 1725–1730.

[34] S. Patwardhan, S. Banerjee, and T. Pedersen, "Senserelate::targetword - a generalized framework for word sense disambiguation," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2005, pp. 73–76.

[35] T. Pedersen and V. Kolhatkar, "Wordnet::senserelate::allwords - a broad coverage word sense tagger that maximimizes semantic relatedness," in *Proceedings of NAACL/HLT 2009*, 2009, pp. 17–20.

[36] I. P. Klapafits and S. Manandhar, "Word sense induction & disambiguation using hierarchical random graphs," in *Proceedings of EMNLP 2010*, 2010, pp. 745–755.

[37] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of SIGDOC 1986*, 1986, pp. 24–26.

[38] R. Mihalcea, "Co-training and self-training for word sense disambiguation," in *Proceedings of CoNLL-2004*, 2004, pp. 33–40.

[39] R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[40] P. Basile, A. Caputo, and G. Semeraro, "An enhanced lesk word sense disambiguation algorithm through a distribution semantic

model," in *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1591–1600.
[41] J. Carroll and D. McCarthy, "Word sense disambiguation using automatically acquired verbal preferences," *Computers and the Humanities*, vol. 34, pp. 109–114, 2000.
[42] D. McCarthy and J. Carroll, "Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences," *Computational Linguistics*, vol. 29, no. 4, pp. 639–654, 2003.
[43] J. F. Cai, W. S. Lee, and Y. W. Teh, "Improving word sense disambiguation using topic features," in *Proceedings of EMNLP-CoNLL 2007*, 2007, pp. 1015–1023.
[44] J. Boyd-Graber, D. Blei, and X. Zhu, "A topic model for word sense disambiguation," in *Proceedings of EMNLP-CoNLL 2007*, 2007, pp. 1024–1033.
[45] L. Li, B. Roth, and C. Sporleder, "Topic models for word sense disambiguation and token-based idiom detection," in *Proceedings of the 48th ACL*, 2010, pp. 1138–1147.
[46] D. O. Séaghdha, "Latent variable models of selectional preference," in *Proceedings of the 48th ACL*, 2010, pp. 435–444.
[47] A. Ritter, Mausam, and O. Etzioni, "A latent dirichlet allocation method for selectional preferences," in *Proceedings of the 48th ACL*, 2010, pp. 424–434.
[48] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.
[49] T. Griffiths and M. Steyvers, "A probabilistic approach to semantic representation," in *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002, pp. 381–386.
[50] T. L. Griffiths and M. Steyvers, "Prediction and semantic association," *Advances in Neural Information Processing Systems 15*, pp. 11–18, 2003. [Online]. Available: http://books.nips.cc/papers/files/nips15/CS02.pdf
[51] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences 101 (Suppl. 1)*, 2004, pp. 5228–5235.
[52] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.
[53] R. Sinha and R. Mihalcea, "Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity," in *Proceedings of the International Conference on Semantic Computing*, ser. ICSC '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 363–369. [Online]. Available: http://dx.doi.org/10.1109/ICSC.2007.107

Yoonjung Choi received her bachelor degree in Computer Science and Engineering and her master degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST), and her ph.D degree in Computer Science from University of Pittsburgh. She is currently employed at Samsung Electronics Co. in South Korea. Her research focuses on sentiment analysis, natural language processing, artificial intelligence, data mining, and machine learning.

Janyce Wiebe is Professor of Computer Science and Co-Director of the Intelligent Systems Program at the University of Pittsburgh. Her research with students and colleagues has been in discourse processing, pragmatics, and word-sense disambiguation. A major concentration of her research is "subjectivity analysis", recognizing and interpreting expressions of opinions and sentiments in text, to support NLP applications such as question answering, information extraction, text categorization, and summarization. Her professional roles have included ACL Program Co-Chair, NAACL Program Chair, NAACL Executive Board member, Computational Linguistics and Language Resources and Evaluation Editorial Board member, AAAI Workshop Co-Chair, ACM Special Interest Group on Artificial Intelligence (SIGART) Vice-Chair, and ACM-SIGART/AAAI Doctoral Consortium Chair.

Rada Mihalcea is a Professor in the Department of Computer Science and Engineering at the University of Michigan. Her research interests are in computational linguistics, with a focus on lexical semantics, graph-based algorithms for natural language processing, and multilingual natural language processing. She serves or has served on the editorial boards of the Journals of Computational Linguistics, Language Resources and Evaluations, Natural Language Engineering, Research in Language in Computation, IEEE Transactions on Affective Computing, and Transactions of the Association for Computational Linguistics. She was a program co-chair for the Conference of the Association for Computational Linguistics (2011) and the Conference on Empirical Methods in Natural Language Processing (2009), and a general chair for the Conference of the North American Association for Computational Linguistics (NAACL 2015). She is the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009).