Augmenting English Adjective Senses with Supersenses

Yulia Tsvetkov[*] **Nathan Schneider**[*] **Dirk Hovy**[†] **Archna Bhatia**[*] **Manaal Faruqui**[*] **Chris Dyer**[*]

[*]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
{ytsvetko,nschneid,archna,mfaruqui,cdyer}@cs.cmu.edu

[†]Center for Language Technology, University of Copenhagen, Denmark
dirk@cst.dk

Abstract

We develop a supersense taxonomy for adjectives, based on that of GermaNet, and apply it to English adjectives in WordNet using human annotation and supervised classification. Results show that accuracy for automatic adjective *type* classification is high, but *synsets* are considerably more difficult to classify, even for trained human annotators. We release the manually annotated data, the classifier, and the induced supersense labeling of 12,304 WordNet adjective synsets.

Keywords: adjective supersenses, lexical semantics, semantic taxonomy induction

1. Introduction

English WordNet (Fellbaum, 1998) offers a fine-grained inventory of semantic senses for adjectives. Like nouns, verbs, and adverbs, these are organized into *synsets* (synonym sets). Unlike nouns and verbs, however, there is no hierarchical taxonomy for adjectives; instead, adjective synsets are organized in **clusters** consisting of a core synset and linked satellite synsets with closely related meanings (Gross and Miller, 1990).[1] Members of these clusters are sometimes linked to nouns or verbs, or to other clusters via "see also" or antonymy links, but there is no systematic organization connecting these clusters. For example, *exasperated* and *cheesed off* are listed as synonyms and *displeased* as closely related, but there is nothing to indicate that these as well as *ashamed* all describe emotional states.

This work presents an approach to eliciting high-level groupings of adjective synsets into a small number of coarse classes. Inspired by WordNet's partitioning of nouns and verbs into semantic field categories now known as **supersenses**[2] (Ciaramita and Altun, 2006; Nastase, 2008), we borrow and adapt to English the top-level adjectival classification scheme[3] from GermaNet (i.e., the German-language WordNet; Hamp and Feldweg, 1997). The 13 English supersense categories appear in Table 1. §2 provides further background on coarse sense taxonomies and motivates the choice of GermaNet's for adjectives.

Due to the large inventory of adjective synsets in English WordNet, full manual annotation by experts would be prohibitively expensive. Because the GermaNet resource is proprietary, we map English adjective senses to these supersenses not by cross-lingual mapping, but with partial human annotation (cf. Schneider et al., 2012):

- With some seed examples, we train a multi-class classifier (§3) to score supersenses for a given adjective *type* (out of context) on the basis of distributional features. The 13 supersenses are ranked for each adjective type in our data according to their classifier scores.
- §4 presents two approaches to aggregating the type annotations into *synset* annotations: the first uses the type-based classifier directly over all lemmas in each synset; the second uses it indirectly to rank choices for a crowdsourcing task, followed by filtering and aggregation.

We publicly release the classifier,[4] the induced supersense labeling of 12,304 WordNet adjective synsets, as well as the raw data (classifier predictions, annotation items, and human judgments).[5]

2. Taxonomy

Several taxonomies of adjectives have been proposed: for English (Dixon, 1982; Raskin and Nirenburg,

[1] There are 18,156 adjective synsets in WordNet (7,463 main synsets and 10,693 "satellite" synsets representing variations on main synsets (mapped with a "similar to" link). The lemmas in these synsets capture 21,479 adjective types, 4,993 of which are polysemous.

[2] Or *lexicographer categories*.

[3] http://www.sfs.uni-tuebingen.de/lsd/adjectives.shtml

[4] https://github.com/ytsvetko/adjective_supersense_classifier

[5] http://www.cs.cmu.edu/~ytsvetko/adj-supersenses.gz

Example words	Supersense	#	Example subclasses
purple, shiny, taut, glittering, smellier, salty, noisy	PERCEPTION	114	color, lightness, taste, smell, sound
compact, gigantic, circular, hollow, adjacent, far	SPATIAL	143	dimension, direction, localization, origin, shape
old, continual, delayed, annual, junior, adult, rapid	TEMPORAL	86	time, age, velocity, periodicity
gliding, flowing, immobile	MOTION	28	motion
creamy, frozen, dense, moist, ripe, closed, metallic, dry	SUBSTANCE	115	consistency, material temperature, physical properties
rainy, balmy, foggy, hazy, humid	WEATHER	25	weather, climate
alive, athletic, muscular, ill, deaf, hungry, female	BODY	100	constitution, affliction, physical sensation, appearance
angry, embarrassed, willing, pleasant, cheerful	FEELING	192	feeling, stimulus
clever, inventive, silly, educated, conscious	MIND	68	intelligence, awareness, knowledge, experience
bossy, deceitful, talkative, tame, organized, adept, popular	BEHAVIOR	178	character, inclination, discipline, skill
affluent, upscale, military, devout, Asian, arctic, rural	SOCIAL	103	stratum, politics, religion, ethnicity, nationality, region
billionth, enough, inexpensive, profitable	QUANTITY	77	number, amount, cost, profit
important, chaotic, affiliated, equal, similar, vague	MISC.	127	order, completeness, validity

Table 1: Adjective taxonomy and example words. Counts in the training data are shown for each category; see §3.1.

1995; Peters and Peters, 2000; Dixon and Aikhenvald, 2004), German (Hundsnurscher and Splett, 1982), Portuguese (Marrafa and Mendes, 2006), and Catalan (Torrent et al., 2012). Following Schneider et al. (2012), we sought a taxonomy with a medium level of granularity such as would facilitate relatively simple annotation tasks, would not be constrained by coverage of an existing lexicon, would be largely applicable across languages, and would facilitate automatic tagging systems (cf. Ciaramita and Altun, 2006; Paaß and Reichartz, 2009; Nastase, 2008) for use in applications like question answering and information extraction.

We chose the taxonomy used for adjectives in GermaNet (Hamp and Feldweg, 1997), which was adapted from Hundsnurscher and Splett (1982). We adopt this taxonomy due to its purely semantic nature and hierarchical structure. There are thirteen coarse semantic classes on the top level, followed by an additional level of finer-grained subcategories. These subcategories facilitate annotation of coarse classes, and can serve a convenient starting point for further research on categorization of adjectives into semantic classes of varying granularities. Table 1 summarizes the top-level supersense categories which we elaborate in this work, along with examples of subcategories and some examples of annotated adjectives.

3. Labeling word types

We build a weakly supervised classifier that labels adjective *types* (irrespective of context). The classifier is trained on a small set of seed examples (§3.1) and uses a feature representation derived in an unsupervised fashion from distributional statistics over large corpora (§3.3). §3.4 shows that this classifier is surprisingly accurate, ranking the correct label near the top of the list for most adjectives.

3.1. Training data

We first collected 1,223 adjectives to use as seed examples for the supersense classes.[6] Some of these are translations of GermaNet examples; others were obtained from the web. 65 of them were assigned to multiple classes; these occur in multiple training instances, so there are 1,294 seed instances in total. 10% of instances from each class (127 total) were randomly sampled as a held-out evaluation set, leaving 1,167 training seeds.

As this would be quite small for training a 13-class classifier, we apply heuristics to expand the data automatically. Candidates are derived from each training seed as follows, and are assigned the same label as the seed:

- **String-based expansion:** We concatenate affixes to the surface form of the adjective to create comparatives and superlatives (*strong* → *strong**er***, *strong**est***) and negations (*manageable* → ***un**manageable*).
- **WordNet-based expansion:** This involves following synonymy, antonymy, and adjective-noun derivational links for all lemmas in all synsets of the word so as to reach additional adjective lemmas.

Although these steps substantially increase the size of the dataset, they introduce considerable noise due to polysemy in WordNet. Therefore, we first train a classifier (with the same method as §3.2) on the initial seeds, and use it to predict a label for the expansion candidate adjectives. If the classifier's predicted label for the expansion candidate agrees with the label of

[6] In the final training set we retain only items on which two independent annotators agree, and for which word vectors were available.

the seed it was derived from, the candidate is added to the training set. This augments the training data by a quarter, resulting in 1,473 training examples (an average of 113 per class). The distribution of supersense examples in the training set, post-expansion, appears in table 1.

3.2. Classification method

We employ the random forest classifier (Breiman, 2001), implemented in the `scikit-learn` toolkit (Pedregosa et al., 2011). A random forest is an ensemble of decision tree classifiers learned from many independent subsamples of the training data. Given an input, each tree classifier assigns a probability to each label; those probabilities are averaged to compute the probability distribution across the ensemble.

This method is particularly suited to our experimental setup: it is known to be effective for multi-label classification and in imbalanced and noisy data scenarios. Moreover, the random forests are immune to overfitting: as the number of trees increases, they produce a limiting value of the generalization error.[7] Thus, no hyperparameter tuning is required. These properties are crucial for our small 13-class training set, in which some of the types are seen in training with more than one label. Posterior probabilities can be obtained from the model in order to rank the classes.

3.3. Features

The sole features in the classifier are vector space word representations built from an unlabeled corpus. The vectors are projections of distributional contexts into 64 dimensions; each dimension is a feature. These features on the one hand effectively capture contextual, semantic properties in their dimensions, and on the other hand are dense enough to keep the model to a small number of parameters.

We use the cross-lingually enriched distributional word vectors constructed in Faruqui and Dyer (2014) using both monolingual word cooccurrence and bilingual word alignment information. These vectors were shown to outperform the traditional word vectors constructed using only monolingual information on a variety of tasks. The released word vectors[8] were trained on the news commentary corpus released by WMT-2011[9] and contain $10,793$ adjective types from our data set.

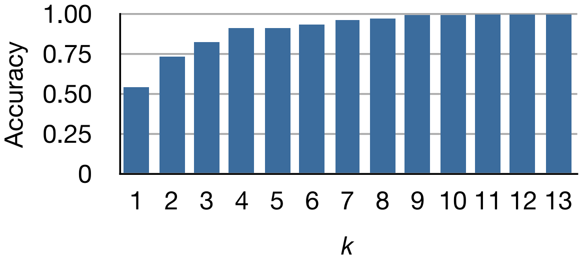


Figure 1: Type classification accuracy for top-*k* semantic class labels. For *k*=4, the classifier accuracy is 91%.

3.4. Type classification evaluation

As noted in §3.1, we train the classifier on 1,473 labeled adjective instances and evaluate on 127 held-out instances. Figure 1 details the classifier's accuracy on the held-out examples. We first rank the classifier's posterior probabilities for each example *w*, and then measure accuracy for each *k*: the prediction for *w* is considered correct if a human-provided supersense label is among the top-*k* items of the posterior.

Thus, with $k = 1$ the accuracy is 54%, which is substantially better than a random baseline of 7% for 13-way classification. With $k = 4$, it is 91%, i.e. we can reliably predict that a correct semantic class is among top-4 labels.

4. Labeling senses

In the previous section, we have shown how a classifier can be trained to label adjective *types*. In the following, we are concerned with labeling a whole *synset* instead. For this, we make active use of the type classifier developed previously.

More concretely, we experiment with two strategies for labeling WordNet adjective *synsets*. The first, §4.1, aggregates predictions from our type-based classifier over the lemmas within each synset. The second, §4.2, elicits human judgments on lemmas-in-context via crowdsourcing, then aggregates these judgments for each synset. §4.3 compares the two techniques.

4.1. Classifier voting

Recall that the classifier described in §3 predicts a supersense class given an adjective *type*. If an adjective is strongly polysemous, this should be reflected in the posterior over supersense labels, but the classifier on its own does not offer a way to decide which instances of the adjective correspond to which supersense.

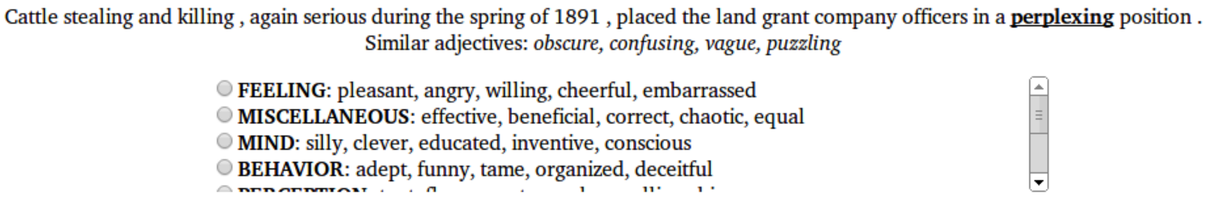
We therefore take advantage of the structure of WordNet, which groups synonymous adjective lemmas into synsets.

The classifier described in §3 predicts a supersense class given an adjective *type*, but most frequent adjec-

[7] See Theorem 1.2 in Breiman (2001) for details.

[8] `http://www.cs.cmu.edu/~mfaruqui/soft.html`

[9] `http://www.statmt.org/wmt11/`



Cattle stealing and killing , again serious during the spring of 1891 , placed the land grant company officers in a **perplexing** position .
Similar adjectives: *obscure, confusing, vague, puzzling*

○ **FEELING**: pleasant, angry, willing, cheerful, embarrassed
○ **MISCELLANEOUS**: effective, beneficial, correct, chaotic, equal
○ **MIND**: silly, clever, educated, inventive, conscious
○ **BEHAVIOR**: adept, funny, tame, organized, deceitful

Figure 2: Screenshot of an item in the crowdsourcing task. The target adjective is highlighted in a SemCor sentence. Supersense choices are ordered according to the posterior from the type-based classifier.

	Novice		Expert		Either
	Acc. (%)	Avg. Rank	Acc. (%)	Avg. Rank	% Acceptable
Classifier: hard voting	33.1	—	40.5	—	48.6
Crowdsourcing	37.8	3.9	43.9	3.1	52.7
Classifier: soft voting	39.9	3.6	44.6	2.9	56.1

Table 2: Comparison of methods for labeling polysemous synsets

tives are polysemous. (In WordNet, nearly a quarter of adjectives are listed under multiple synsets.) We hypothesize that different members of a synset will tend to show different polysemy patterns over supersenses, and that combining these will reveal the supersense most closely associated with the common semantics of the synset in question. For instance, though *sweet* is ambiguous between a taste (PERCEPTION) and a personality trait (BEHAVIOR), where it appears in a cluster with *lovely* and *endearing*, the PERCEPTION reading can be ruled out.

This hypothesis motivates a simple *voting* regimen using the type-based classifier. That is, for each synset, we ask our classifier to predict a supersense label for each lemma in that synset, and let these lemmas "vote" on the overall supersense label for the synset. We consider two voting schemes: **hard** voting, in which each lemma votes for the top classifier prediction;[10] and **soft** voting, in which the classifier posteriors are averaged over the lemmas in each synset. The supersense getting the most (hard or soft) votes from lemmas is predicted as the label for the synset.

4.2. Crowdsourcing

As an inexpensive source of human labels for adjective senses, we turned to crowdsourcing with the Amazon Mechanical Turk (AMT) platform. AMT allows web-based tasks—known as "human intelligence tasks" (**HIT**s)—to be published to users (**workers**) who have the opportunity to complete them for a small fee. Though the workers are not necessarily qualified to make advanced linguistic judgments, researchers have found ways to construct simple language-oriented tasks and to obtain accurate results quickly and cheaply, provided sufficient quality control mechanisms are in place (Snow et al., 2008; Munro et al., 2010).

A synopsis of our methodology is follows:[11]

1. Workers are shown SemCor (Miller et al., 1993) sentences with an adjective token highlighted in context[12] and instructed to choose the most contextually relevant label for the adjective from a list of supersense presorted according to the type-based classifier (§3).[13] An item from this task is visualized in figure 2. We obtain 3–5 judgments for 5077 sentences, one per adjective lemma represented in SemCor for which word vectors were available. 10 sentences were randomly assigned to each HIT, for which the worker was offered 7 cents. The HITs were completed over about 3 days for a total cost of $170.

2. Responses for each lemma are aggregated with the MACE tool[14] (Hovy et al., 2013), which takes annotator-specific patterns into account to better determine the true label. Because workers are untrained and incentivized to complete HITs as quickly as possible, careful quality control and noise tolerance are required even for the simplest of tasks. Obviously problematic users were rejected from the task, and heuristics were used to

[10] Ties are broken randomly.

[11] Space constraints do not allow us to report all details here, but we will document them in the data release.

[12] We decided it would be impractical to show full synset or lemma descriptions directly to workers for labeling, as many of the definitions provided in WordNet may be too technical.

[13] The purpose of the presorting was to make the task faster and less visually overwhelming for the user. Only the top-4 classes were displayed by default (with the others reachable by scrolling).

[14] `http://www.isi.edu/publications/licensed-sw/mace/`

	Gold											
Classifier (soft)	BEHAVIOR	BODY	FEELING	MIND	MISC.	PERCEPTION	QUANTITY	SOCIAL	SPATIAL	SUBSTANCE	TEMPORAL	**total**
BEHAVIOR	**2**		1		3							6
BODY	1	**5**	2			1			1	1		11
FEELING	8	2	**14**		6		1	1		1	2	35
MIND	2			**1**		1						4
MISC.			1		**14**						1	16
PERCEPTION	2	2	1	1	5	**4**						15
QUANTITY					2							2
SOCIAL	1				1			**5**				7
SPATIAL	2	2	2		5		1	3	**16**		1	32
SUBSTANCE			2		1	1			1	**3**		8
TEMPORAL	2		1		2				1		**6**	12
total	20	11	24	2	39	7	2	9	19	5	10	148

Table 3: Confusion matrix for classifier soft voting vs. either gold annotation

filter to 486 high-confidence synset clusters covering 1791 lemmas.

3. For each of the high-confidence clusters, we aggregate predictions over lemmas in the cluster to choose a single supersense label, again using MACE. This results in 438 synset clusters (covering 578 lemmas) labeled with supersenses.

4.3. Synset Evaluation

To evaluate the two methods of sense-level labeling, we sampled 148 multi-lemma adjective clusters from the 438 high-confidence crowdsourced predictions. Two graduate students (both authors of this paper) manually labeled these clusters, consulting their WordNet descriptions (list of lemmas, synset definitions, curated example phrases). One annotator was well acquainted with the adjective supersense scheme, while the other received a minimal amount of training in the task (comparable to that of the workers in crowdsourcing). Their inter-annotator agreement rate was 55%, with a Cohen's κ of 0.47 ("moderate"). This reinforces the difficulty of choosing a single class for many of the senses, despite their intuitive applicability to prototypical members. There were clear patterns of disagreement (e.g., one annotator applied the FEELING category much more liberally) that would likely have been resolved with more training/discussion of the annotation standard.

Table 2 uses these annotations as a gold standard for comparing the fully automated method of §4.1 vs. the crowdsourcing-based method of §4.2. For each of the annotators, it measures accuracy as well as the average rank of the annotator's label in the ordering implied by the automatic scheme (where applicable). The "Either" column represents a more lenient evaluation method (if the Novice and Expert disagreed, both their labels are considered acceptable).

The apparent trends are that (a) the automatic predictions are closer to the Expert annotations than they are to the Novice annotations, and (b) the best method is soft voting with the classifier. Note that we have many more predictions from the classifier (12,304 clusters with word vectors synset clusters) than from the crowdsourcing exercise (just 438 after filtering steps), so extrapolating these results to the classifier's other predictions, this casts doubt on the utility of further crowdsourcing experiments to try to expand coverage.

The confusion matrix between the best automatic method and the gold annotations (table 3) illustrates that recurring pairs of classes are easily confusable: aspects of FEELING and BEHAVIOR may simultaneously be captured in adjectives such as *calm*, *mad*, and *joyful*. MISCELLANEOUS is the least semantically cohesive class, which likely accounts for its high error rate.

5. Conclusion

We semi-automatically augment WordNet senses with high-level semantic classes (supersenses) for adjectives. Our techniques include manual annotation, crowdsourcing, and training supervised classifiers. The resulting 12,304 supersense-labeled synsets are publicly released along with supplementary data. These resources have already been found useful in a downstream task: Tsvetkov et al. (2014) used adjective type supersenses to improve the classification accuracy of adjective-noun metaphoric and literal pairs in mono- and cross-lingual settings.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533.

References

Breiman, Leo (2001). Random forests. *Machine Learning*, 45(1):5–32.

Ciaramita, Massimiliano and Altun, Yasemin (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602. Association for Computational Linguistics.

Dixon, Robert M. W. (1982). Where have all the adjectives gone? In Dixon, Robert M. W., editor, *Where Have All the Adjectives Gone?: And Other Essays in Semantics and Syntax*, pages 1–62. Mouton.

Dixon, Robert M. W. and Aikhenvald, Aleksandra (2004). *Adjective classes*. Oxford University Press.

Faruqui, Manaal and Dyer, Chris (2014). Improving vector space word representations using multilingual correlation. In *Proc. of EACL*. Association for Computational Linguistics.

Fellbaum, Christiane, editor (1998). *WordNet: an electronic lexical database*. MIT Press.

Gross, Derek and Miller, Katherine J. (1990). Adjectives in WordNet. *International Journal of Lexicography*, 3(4):265–277.

Hamp, Birgit and Feldweg, Helmut (1997). GermaNet - a lexical-semantic net for German. In *Proc. of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Association for Computational Linguistics.

Hovy, Dirk, Berg-Kirkpatrick, Taylor, Vaswani, Ashish, and Hovy, Eduard (2013). Learning whom to trust with MACE. In *Proc. of NAACL-HLT*, pages 1120–1130. Association for Computational Linguistics.

Hundsnurscher, Franz and Splett, Jochen (1982). *Semantik der Adjektive des Deutschen*. 3137. Westdeutscher Verlag.

Marrafa, Palmira and Mendes, Sara (2006). Modeling adjectives in computational relational lexica. In *Proc. of COLING*, pages 555–562. Association for Computational Linguistics.

Miller, George A., Leacock, Claudia, Tengi, Randee, and Bunker, Ross T. (1993). A semantic concordance. In *Proc. of HLT*, pages 303–308. Association for Computational Linguistics.

Munro, Robert, Bethard, Steven, Kuperman, Victor, Lai, Vicky Tzuyin, Melnick, Robin, Potts, Christopher, Schnoebelen, Tyler, and Tily, Harry (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.

Nastase, Vivi (2008). Unsupervised all-words word sense disambiguation with grammatical dependencies. In *Proc. of IJCNLP*, pages 7–12. Asian Federation of Natural Language Processing.

Paaß, Gerhard and Reichartz, Frank (2009). Exploiting semantic constraints for estimating supersenses with CRFs. In *Proc. of SDM*. Society for Industrial and Applied Mathematics.

Pedregosa, Fabian, Varoquaux, Gael, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, M., and Duchesnay, Edouard (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, Ivonne and Peters, Wim (2000). The treatment of adjectives in SIMPLE: Theoretical observations. In *LREC*. European Language Resources Association.

Raskin, Victor and Nirenburg, Sergei (1995). Lexical semantics of adjectives. Technical Report MCCS-95-288, New Mexico State University Computing Research Laboratory, Las Cruces, NM. `http://web.ics.purdue.edu/~vraskin/adjective.pdf`.

Schneider, Nathan, Mohit, Behrang, Oflazer, Kemal, and Smith, Noah A. (2012). Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proc. of ACL*, pages 253–258. Association for Computational Linguistics.

Snow, Rion, O'Connor, Brendan, Jurafsky, Dan, and Ng, Andrew Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, pages 254–263. Association for Computational Linguistics.

Torrent, Gemma Boleda, im Walde, Sabine Schulte, and Badia, Toni (2012). Modeling regular polysemy: A study on the semantic classification of Catalan adjectives. *Computational Linguistics*, 38(3):575–616.

Tsvetkov, Yulia, Boytsov, Leonid, Gershman, Anatole, Nyberg, Eric, and Dyer, Chris (2014). Metaphor detection with cross-lingual model transfer. In *Proc. of ACL*. Association for Computational Linguistics.