Regular polysemy: A distributional model

Gemma Boleda
Dept. of Linguistics
University of Texas at Austin
gemma.boleda@upf.edu

Sebastian Padó
ICL
University of Heidelberg
pado@cl.uni-heidelberg.de

Jason Utt
IMS
University of Stuttgart
uttjn@ims.uni-stuttgart.de

Abstract

Many types of polysemy are not word specific, but are instances of general sense alternations such as ANIMAL-FOOD. Despite their pervasiveness, regular alternations have been mostly ignored in empirical computational semantics. This paper presents (a) a general framework which grounds sense alternations in corpus data, generalizes them above individual words, and allows the prediction of alternations for new words; and (b) a concrete unsupervised implementation of the framework, the Centroid Attribute Model. We evaluate this model against a set of 2,400 ambiguous words and demonstrate that it outperforms two baselines.

1 Introduction

One of the biggest challenges in computational semantics is the fact that many words are *polysemous*. For instance, *lamb* can refer to an animal (as in *The lamb squeezed through the gap*) or to a food item (as in *Sue had lamb for lunch*). Polysemy is pervasive in human language and is a problem in almost all applications of NLP, ranging from Machine Translation (as word senses can translate differently) to Textual Entailment (as most lexical entailments are sense-specific).

The field has thus devoted a large amount of effort to the representation and modeling of word senses. The arguably most prominent effort is Word Sense Disambiguation, WSD (Navigli, 2009), an *in-vitro* task whose goal is to identify which, of a set of predefined senses, is the one used in a given context.

In work on WSD and other tasks related to polysemy, such as word sense induction, sense alternations are treated as *word-specific*. As a result, a model for the meaning of *lamb* that accounts for the relation between the animal and food senses cannot predict that the same relation holds between instances of *chicken* or *salmon* in the same type of contexts.

A large number of studies in linguistics and cognitive science show evidence that there are regularities in the way words vary in their meaning (Apresjan, 1974; Lakoff and Johnson, 1980; Copestake and Briscoe, 1995; Pustejovsky, 1995; Gentner et al., 2001; Murphy, 2002), due to general analogical processes such as regular polysemy, metonymy and metaphor. Most work in theoretical linguistics has focused on *regular*, *systematic*, or *logical* polysemy, which accounts for alternations like ANIMAL-FOOD. Sense alternations also arise from metaphorical use of words, as *dark* in *dark glass-dark mood*, and also from metonymy when, for instance, using the name of a place for a representative (as in *Germany signed the treatise*). Disregarding this evidence is empirically inadequate and leads to the well-known *lexical bottleneck* of current word sense models, which have serious problems in achieving high coverage (Navigli, 2009).

We believe that empirical computational semantics could profit from a model of polysemy[1] which (a) is applicable across individual words, and thus capable of capturing general patterns and generalizing to new

[1] Our work is mostly inspired in research on regular polysemy. However, given the fuzzy nature of "regularity" in meaning variation, we extend the focus of our attention to include other types of analogical sense construction processes.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 151–160,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

words, and (b) is induced in an unsupervised fashion from corpus data. This is a long-term goal with many unsolved subproblems.

The current paper presents two contributions towards this goal. First, since we are working on a relatively unexplored area, we introduce a formal framework that can encompass different approaches (Section 2). Second, we implement a concrete instantiation of this framework, the unsupervised Centroid Attribute Model (Section 3), and evaluate it on a new task, namely, to detect which of a set of words instantiate a given type of polysemy (Sections 4 and 5). We finish with some conclusions and future work (Section 7).

2 Formal framework

In addition to introducing formal definitions for terms commonly found in the literature, our framework provides novel terminology to deal with regular polysemy in a general fashion (cf. Table 1; capital letters designate sets and small letters elements of sets).[2]

For a lemma l like *lamb*, we want to know how well a **meta alternation** (such as ANIMAL-FOOD) explains a pair of its senses (such as the animal and food senses of *lamb*).[3] This is formalized through the function score, which maps a meta alternation and two senses onto a score. As an example, let $\textit{lamb}_{\mathtt{anm}}$ denote the ANIMAL sense of *lamb*, $\textit{lamb}_{\mathtt{fod}}$ the FOOD sense, and $\textit{lamb}_{\mathtt{hum}}$ the PERSON sense. Then, an appropriate model of meta alternations should predict that $\mathrm{score}(\mathtt{animal}, \mathtt{food}, \textit{lamb}_{\mathtt{anm}}, \textit{lamb}_{\mathtt{fod}})$ is greater than $\mathrm{score}(\mathtt{animal}, \mathtt{food}, \textit{lamb}_{\mathtt{anm}}, \textit{lamb}_{\mathtt{hum}})$.

Meta alternations are defined as unordered pairs of **meta senses**, or cross-word senses like ANIMAL. The meta senses M can be defined a priori or induced from data. They are equivalence classes of senses to which they are linked through the function meta. A sense s *instantiates* a meta sense m iff $\mathrm{meta}(s) = m$. Functions inst and sns allow us to define meta senses and lemma-specific senses in terms of actual instances, or occurrences of words in context.

L	set of lemmas
I_L	set of (lemma-wise) instances
S_L	set of (lemma-wise) senses
$\mathrm{inst}\colon L \rightarrow \wp(I_L)$	mapping lemma → instances
$\mathrm{sns}\colon L \rightarrow \wp(S_L)$	mapping lemma → senses
M	set of meta senses
$\mathrm{meta}\colon S_L \rightarrow M$	mapping senses → meta senses
$A \subseteq M \times M$	set of meta alternations (MAs)
$\mathfrak{A}$	set of MA representations
$\mathrm{score}\colon A \times S_L^2 \rightarrow \mathbb{R}$	scoring function for MAs
$\mathrm{rep_A}\colon A \rightarrow \mathfrak{A}$	MA representation function
$\mathrm{comp}\colon \mathfrak{A} \times S_L^2 \rightarrow \mathbb{R}$	compatibility function

Table 1: Notation and signatures for our framework.

We decompose the score function into two parts: a *representation* function $\mathrm{rep_A}$ that maps a meta alternation into some suitable representation for meta alternations, $\mathfrak{A}$, and a *compatibility* function comp that compares the relation between the senses of a word to the meta alternation's representation. Thus, $\mathrm{comp} \circ \mathrm{rep_A} = \mathrm{score}$.

3 The Centroid Attribute Model

The *Centroid Attribute Model (CAM)* is a simple instantiation of the framework defined in Section 2, designed with two primary goals in mind. First, it is a data-driven model. Second, it does not require any manual sense disambiguation, a notorious bottleneck.

To achieve the first goal, CAM uses a distributional approach. It represents the relevant entities as *co-occurrence vectors* that can be acquired from a large corpus (Turney and Pantel, 2010). To achieve the second goal, CAM represents meta senses using *monosemous* words only, that is, words whose senses all correspond to one meta sense. [4] Examples are *cattle* and *robin* for the meta sense ANIMAL. We define the vector for a meta sense as the *centroid* (average vector) of the monosemous words instantiating it. In turn, meta alternations are represented by the centroids of their meta senses' vectors.

This strategy is not applicable to test lemmas, which instantiate some meta alternation and are by definition ambiguous. To deal with these without

[2] We re-use inst as a function that returns the set of instances for a sense: $S_L \rightarrow \wp(I_L)$ and assume that senses partition lemmas' instances: $\forall l : \mathrm{inst}(l) = \bigcup_{s \in \mathrm{sns}(l)} \mathrm{inst}(s)$.

[3] Consistent with the theoretical literature, this paper focuses on two-way polysemy. See Section 7 for further discussion.

[4] 10.8% of noun types in the corpus we use are monosemous and 2.3% are disemous, while, on a token level, 23.3% are monosemous and 20.2% disemous.

$\mathrm{vec}_I : I_L \to \mathbb{R}^k$	instance vector computation
$\mathfrak{C} : \mathbb{R}^{k \times m} \to \mathbb{R}^k$	centroid computation
$\mathrm{vec}_L : L \to \mathbb{R}^k$	lemma (type) vector computation
$\mathrm{rep}_\mathrm{M} : M \to \mathbb{R}^k$	meta sense representation

Table 3: Additional notation and signatures for CAM

explicit sense disambiguation, CAM represents lemmas by their type vectors, i.e., the *centroid* of their instances, and compares their vectors (*attributes*) to those of the meta alternation – hence the name.

CoreLex: A Semantic Inventory. CAM uses CoreLex (Buitelaar, 1998) as its meta sense inventory. CoreLex is a lexical resource that was designed specifically for the study of polysemy. It builds on WordNet (Fellbaum, 1998), whose sense distinctions are too fine-grained to describe general sense alternations. CoreLex defines a layer of abstraction above WordNet consisting of 39 *basic types*, coarse-grained ontological classes (Table 2). These classes are linked to one or more Wordnet *anchor nodes*, which define a mapping from WordNet synsets onto basic types: A synset s maps onto a basic type b if b has an anchor node that dominates s and there is no other anchor node on the path from b and s.[5]

We adopt the WordNet synsets as S, the set of senses, and the CoreLex basic types as our set of meta senses M. The meta function (mapping word senses onto meta senses) is given directly by the anchor mapping defined in the previous paragraph. This means that the set of meta alternations is given by the set of pairs of basic types. Although basic types do not perfectly model meta senses, they constitute an approximation that allows us to model many prominent alternations such as ANIMAL-FOOD.

Vectors for Meta Senses and Alternations. All representations used by CAM are co-occurrence vectors in R^k (i.e., $\mathfrak{A} := R^k$). Table 3 lists new concepts that CAM introduces to manipulate vector representations. vec_I returns a vector for a lemma instance, vec_L a (type) vector for a lemma, and $\mathfrak{C}$ the centroid of a set of vectors.

We leave vec_I and $\mathfrak{C}$ unspecified: we will experiment with these functions in Section 4. CAM does fix the definitions for vec_L and rep_A. First, vec_L defines a lemma's vector as the centroid of its instances:

$$\mathrm{vec}_L(l) = \mathfrak{C}\{\mathrm{vec}_I(i) \mid i \in \mathrm{inst}(l)\} \tag{1}$$

Before defining rep_A, we specify a function rep_M that computes vector representations for meta senses m. In CAM, this vector is defined as the centroid of the vectors for all monosemous lemmas whose WordNet sense maps onto m:

$$\mathrm{rep}_\mathrm{M}(m) = \mathfrak{C}\{\mathrm{vec}_L(l) \mid \mathrm{meta}(\mathrm{sns}(l)) = \{m\}\} \tag{2}$$

Now, rep_A can be defined simply as the centroid of the meta senses instantiating a:

$$\mathrm{rep}_\mathrm{A}(m_1, m_2) = \mathfrak{C}\{\mathrm{rep}_\mathrm{M}(m_1), \mathrm{rep}_\mathrm{M}(m_2)\} \tag{3}$$

Predicting Meta Alternations. The final component of CAM is an instantiation of comp (cf. Table 1), i.e., the degree to which a sense pair (s_1, s_2) matches a meta alternation a. Since CAM does not represent these senses separately, we define comp as

$$\begin{aligned}\mathrm{comp}(a, s_1, s_2) = &\,\mathrm{sim}(a, \mathrm{vec}_L(l)) \\ &\text{so that } \{s_1, s_2\} = \mathrm{sns}(l)\end{aligned} \tag{4}$$

The complete model, score, can now be stated as:

$$\begin{aligned}\mathrm{score}(m, m', s, s') = &\,\mathrm{sim}(\mathrm{rep}_\mathrm{A}(m, m'), \mathrm{vec}_L(l)) \\ &\text{so that } \{s, s'\} = \mathrm{sns}(l)\end{aligned} \tag{5}$$

CAM thus assesses how well a meta alternation $a = (m, m')$ explains a lemma l by comparing the centroid of the meta senses m, m' to l's centroid.

Discussion. The central feature of CAM is that it avoids word sense disambiguation, although it still relies on a predefined sense inventory (WordNet, through CoreLex). Our use of monosemous words to represent meta senses and meta alternations goes beyond previous work which uses monosemous words to disambiguate polysemous words in context (Izquierdo et al., 2009; Navigli and Velardi, 2005).

Because of its focus on avoiding disambiguation, CAM simplifies the representation of meta alternations and polysemous words to single centroid vectors. In the future, we plan to induce word senses (Schütze, 1998; Pantel and Lin, 2002; Reisinger and Mooney, 2010), which will allow for more flexible and realistic models.

[5] This is necessary because some classes have non-disjoint anchor nodes: e.g., ANIMALs are a subset of LIVING BEINGs.

abs	ABSTRACTION	ent	ENTITY	loc	LOCATION	prt	PART
act	ACT	evt	EVENT	log	GEO. LOCATION	psy	PSYCHOL. FEATURE
agt	AGENT	fod	FOOD	mea	MEASURE	qud	DEFINITE QUANTITY
anm	ANIMAL	frm	FORM	mic	MICROORGANISM	qui	INDEFINITE QUANTITY
art	ARTIFACT	grb	BIOLOG. GROUP	nat	NATURAL BODY	rel	RELATION
atr	ATTRIBUTE	grp	GROUPING	phm	PHENOMENON	spc	SPACE
cel	CELL	grs	SOCIAL GROUP	pho	PHYSICAL OBJECT	sta	STATE
chm	CHEMICAL	hum	HUMAN	plt	PLANT	sub	SUBSTANCE
com	COMMUNICATION	lfr	LIVING BEING	pos	POSSESSION	tme	TIME
con	CONSEQUENCE	lme	LINEAR MEASURE	pro	PROCESS	pro	PROCESS

Table 2: CoreLex's basic types with their corresponding WordNet anchors. CAM adopts these as meta senses.

4 Evaluation

We test CAM on the task of identifying which lemmas of a given set instantiate a specific meta alternation. We let the model rank the lemmas through the score function (cf. Table (1) and Eq. (5)) and evaluate the ranked list using Average Precision. While an alternative would be to rank meta alternations for a given polysemous lemma, the method chosen here has the benefit of providing data on the performance of individual meta senses and meta alternations.

4.1 Data

All modeling and data extraction was carried out on the written part of the British National Corpus (BNC; Burnage and Dunlop (1992)) parsed with the C&C tools (Clark and Curran, 2007). [6]

For the evaluation, we focus on *disemous words*, words which instantiate exactly two meta senses according to WordNet. For each meta alternation (m, m'), we evaluate CAM on a set of disemous *targets* (lemmas that instantiate (m, m')) and disemous *distractors* (lemmas that do not). We define three types of distractors: (1) distractors sharing m with the targets (but not m'), (2) distractors sharing m' with the targets (but not m), and (3) distractors sharing neither. In this way, we ensure that CAM cannot obtain good results by merely modeling the similarity of targets to either m or m', which would rather be a coarse-grained word sense modeling task.

To ensure that we have enough data, we evaluate CAM on all meta alternations with at least ten targets that occur at least 50 times in the corpus, discarding nouns that have fewer than 3 characters or contain non-alphabetical characters. The distractors are chosen so that they match targets in frequency. This leaves us with 60 meta alternations, shown in Table 5. For each meta alternation, we randomly select 40 lemmas as experimental items (10 targets and 10 distractors of each type) so that a total of 2,400 lemmas is used in the evaluation.[7] Table 4 shows four targets and their distractors for the meta alternation ANIMAL-FOOD.[8]

4.2 Evaluation Measure and Baselines

To measure success on this task, we use Average Precision (AP), an evaluation measure from IR that reaches its maximum value of 1 when all correct items are ranked at the top (Manning et al., 2008). It interpolates the precision values of the top-n prediction lists for all positions n in the list that contain a target. Let $T = \langle q_1, \ldots, q_m \rangle$ be the list of targets, and let $P = \langle p_1, \ldots, p_n \rangle$ be the list of predictions as ranked by the model. Let $I(x_i) = 1$ if $p_i \in T$, and zero otherwise. Then $AP(P, T) = \frac{1}{m} \sum_{i=1}^{m} I(x_i) \frac{\sum_{j=1}^{i} I(x_i)}{i}$. AP measures the quality of the ranked list for a single meta alternation. The overall quality of a model is given by Mean Average Precision (MAP), the mean of the AP values for all meta alternations.

We consider two baselines: (1) A *random baseline* that ranks all lemmas in random order. This baseline is the same for all meta alternations, since the distribution is identical. We estimate it by sampling. (2) A meta alternation-specific *frequency baseline* which orders the lemmas by their corpus frequencies. This

[6] The C&C tools were able to reliably parse about 40M words.

[7] Dataset available at http://www.nlpado.de/~sebastian/data.shtml.

[8] Note that this experimental design avoids any overlap between the words used to construct sense vectors (one meta sense) and the words used in the evaluation (two meta senses).

Targets	Distractors with meta sense anm	Distractors with meta sense fod	Random distractors
carp	*amphibian* (anm-art)	*mousse* (art-fod)	*appropriation* (act-mea)
duckling	*ape* (anm-hum)	*parsley* (fod-plt)	*scissors* (act-art)
eel	*leopard* (anm-sub)	*pickle* (fod-sta)	*showman* (agt-hum)
hare	*lizard* (anm-hum)	*pork* (fod-mea)	*upholstery* (act-art)

Table 4: Sample of experimental items for the meta alternation anm-fod. (Abbreviations are listed in Table 2.)

baseline uses the intuition that frequent words will tend to exhibit more typical alternations.

4.3 Model Parameters

There are four more parameters to set.

Definition of vector space. We instantiate the vec_I function in three ways. All three are based on dependency-parsed spaces, following our intuition that topical similarity as provided by window-based spaces is insufficient for this task. The functions differ in the definition of the space's dimensions, incorporating different assumptions about distributional differences among meta alternations.

The first option, gram, uses grammatical paths of lengths 1 to 3 as dimensions and thus characterizes lemmas and meta senses in terms of their grammatical context (Schulte im Walde, 2006), with a total of 2,528 paths. The second option, lex, uses words as dimensions, treating the dependency parse as a co-occurrence filter (Padó and Lapata, 2007), and captures topical distinctions. The third option, gramlex, uses lexicalized dependency paths like *obj–see* to mirror more fine-grained semantic properties (Grefenstette, 1994). Both lex and gramlex use the 10,000 most frequent items in the corpus.

Vector elements. We use "raw" corpus co-occurrence frequencies as well as log-likelihood-transformed counts (Lowe, 2001) as elements of the co-occurrence vectors.

Definition of centroid computation. There are three centroid computations in CAM: to combine instances into lemma (type) vectors (function vec_L in Eq. (1)); to combine lemma vectors into meta sense vectors (function rep_M in Eq. (2)); and to combine meta sense vectors into meta alternation vectors (function rep_A in Eq. (3)).

For vec_L, the obvious definition of the centroid function is as a *micro-average*, that is, a simple average over all instances. For rep_M and rep_A, there is a design choice: The centroid can be computed by micro-averaging as well, which assigns a larger weight to more frequent lemmas (rep_M) or meta senses (rep_A). Alternatively, it can be computed by *macro-averaging*, that is, by normalizing the individual vectors before averaging. This gives equal weight to the each lemma or meta sense, respectively. Macro-averaging in rep_A thus assumes that senses are equally distributed, which is an oversimplification, as word senses are known to present skewed distributions (McCarthy et al., 2004) and vectors for words with a predominant sense will be similar to the dominant meta sense vector. Micro-averaging partially models sense skewedness under the assumption that word frequency correlates with sense frequency.

Similarity measure. As the vector similarity measure in Eq. (5), we use the standard cosine similarity (Lee, 1999). It ranges between -1 and 1, with 1 denoting maximum similarity. In the current model where the vectors do not contain negative counts, the range is $[0; 1]$.

5 Results

Effect of Parameters The four parameters of Section 4.3 (three space types, macro-/micro-averaging for rep_M and rep_A, and log-likelihood transformation) correspond to 24 instantiations of CAM.

Figure 1 shows the influence of the four parameters. The only significant difference is tied to the use of lexicalized vector spaces (gramlex / lex are better than gram). The statistical significance of this difference was verified by a t-test ($p < 0.01$). This indicates that meta alternations can be characterized better through fine-grained semantic distinctions than by syntactic ones.

The choice of micro- vs. macro-average does not have a clear effect, and the large variation observed in Figure 1 suggests that the best setup is dependent on the specific meta sense or meta alternation being

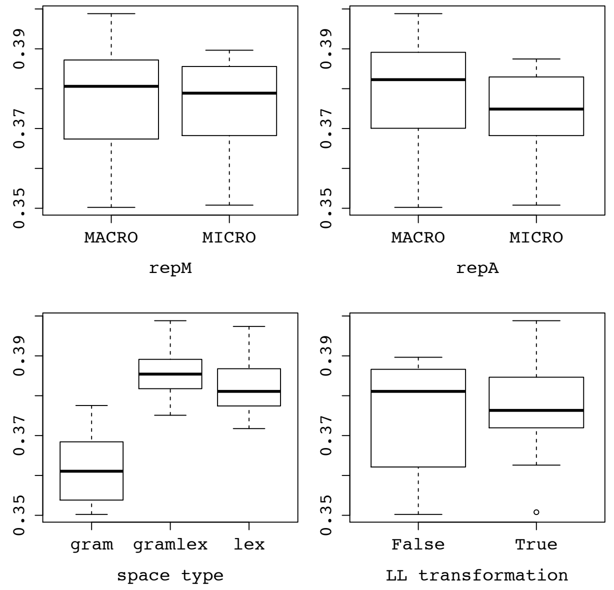

Figure 1: Effect of model parameters on performance. A data point is the mean AP (MAP) across all meta alternations for a specific setting.

modeled. Focusing on meta alternations, whether the two intervening meta senses should be balanced or not can be expected to depend on the frequencies of the concepts denoted by each meta sense, which vary for each case. Indeed, for AGENT-HUMAN, the alternation which most benefits from the micro-averaging setting, the targets are much more similar to the HUMAN meta sense (which is approximately 8 times as frequent as AGENT) than to the AGENT meta sense. The latter contains anything that can have an effect on something, e.g. *emulsifier, force, valium*. The targets for AGENT-HUMAN, in contrast, contain words such as *engineer, manipulator, operative*, which alternate between an agentive role played by a person and the person herself.

While lacking in clear improvement, log-likelihood transformation tends to reduce variance, consistent with the effect previously found in selectional preference modeling (Erk et al., 2010).

Overall Performance Although the performance of the CAM models is still far from perfect, all 24 models obtain MAP scores of 0.35 or above, while the random baseline is at 0.313, and the overall frequency baseline at 0.291. Thus, all models consistently outperform both baselines. A bootstrap resampling test (Efron and Tibshirani, 1994) confirmed that the difference to the frequency baseline is significant at $p < 0.01$ for all 24 models. The difference to the random baseline is significant at $p < 0.01$ for 23 models and at $p < 0.05$ for the remaining model. This shows that the models capture the meta alternations to some extent. The best model uses macro-averaging for rep_M and rep_A in a log-likelihood transformed `gramlex` space and achieves a MAP of 0.399.

Table 5 breaks down the performance of the best CAM model by meta alternation. It shows an encouraging picture: CAM outperforms the frequency baseline for 49 of the 60 meta alternations and both baselines for 44 (73.3%) of all alternations. The performance shows a high degree of variance, however, ranging from 0.22 to 0.71.

Analysis by Meta Alternation Coherence Meta alternations vary greatly in their difficulty. Since CAM is an attribute similarity-based approach, we expect it to perform better on the alternations whose meta senses are ontologically more similar. We next test this hypothesis.

Let $D_{m_i} = \{d_{ij}\}$ be the set of distractors for the targets $T = \{t_j\}$ that share the meta sense m_i, and $D_R = \{d_{3j}\}$ the set of random distractors. We define the coherence κ of an alternation a of meta senses m_1, m_2 as the mean (∅) difference between the similarity of each target vector to a and the similarity of the corresponding distractors to a, or formally $\kappa(a) = \emptyset\ \mathrm{sim}(\mathrm{rep}_\mathrm{A}(m_1, m_2), \mathrm{vec}_L(t_j)) - \mathrm{sim}(\mathrm{rep}_\mathrm{A}(m_1, m_2), \mathrm{vec}_L(d_{ij}))$, for $1 \leq i \leq 3$ and $1 \leq j \leq 10$. That is, κ measures how much more similar, on average, the meta alternation vector is to the target vectors than to the distractor vectors. For a meta alternation with a higher κ, the targets should be easier to distinguish from the distractors.

Figure 2 plots AP by κ for all meta alternations. As we expect from the definition of κ, AP is strongly correlated with κ. However, there is a marked Y shape, i.e., a divergence in behavior between high-κ and mid-AP alternations (upper right corner) and mid-κ and high-AP alternations (upper left corner).

In the first case, meta alternations perform worse than expected, and we find that this typically points to missing senses, that is, problems in the underlying lexical resource (WordNet, via CoreLex). For instance, the FOOD-PLANT distractor *almond* is given

grs-psy	0.709	**com-evt**	0.501	**art-com**	0.400	**atr-com**	0.361	art-frm	0.286
pro-sta	0.678	**art-grs**	0.498	**act-pos**	0.396	**atr-sta**	0.361	**act-hum**	0.281
fod-plt	0.645	**hum-psy**	0.486	**phm-sta**	0.388	**act-phm**	0.339	art-fod	0.280
psy-sta	0.630	**hum-nat**	0.456	**atr-psy**	0.384	**anm-art**	0.335	**grs-hum**	0.272
hum-prt	0.602	**anm-hum**	0.448	**fod-hum**	0.383	**art-atr**	0.333	act-art	0.267
grp-psy	0.574	**com-psy**	0.443	**plt-sub**	0.383	**act-psy**	0.333	art-grp	0.258
grs-log	0.573	**act-grs**	0.441	**act-com**	0.382	**agt-hum**	0.319	art-nat	0.248
act-evt	0.539	**atr-rel**	0.440	**grp-grs**	0.379	**art-evt**	0.314	act-atr	0.246
evt-psy	0.526	**art-qui**	0.433	**art-psy**	0.373	**atr-evt**	0.312	art-hum	0.240
act-tme	0.523	**act-sta**	0.413	**art-prt**	0.364	**art-sta**	0.302	art-loc	0.238
art-pho	0.520	**art-sub**	0.412	**evt-sta**	0.364	act-grp	0.296	art-pos	0.228
act-pro	0.513	**art-log**	0.407	**anm-fod**	0.361	**com-hum**	0.292	com-sta	0.219

Table 5: Meta alternations and their average precision values for the task. The random baseline performs at 0.313 while the frequency baseline ranges from 0.255 to 0.369 with a mean of 0.291. Alternations for which the model outperforms the frequency baseline are in boldface (mean AP: 0.399, standard deviation: 0.119).

grs-psy	*democracy, faculty, humanism, regime,*
pro-sta	*bondage, dehydration, erosion,urbanization*
psy-sta	*anaemia,delight, pathology, sensibility*
hum-prt	*bum, contractor, peter, subordinate*
grp-psy	*category, collectivism, socialism, underworld*

Table 6: Sample targets for meta alternations with high AP and mid-coherence values.

a PLANT sense by WordNet, but no FOOD sense. In the case of SOCIAL GROUP-GEOGRAPHICAL LOCATION, distractors *laboratory* and *province* are missing SOCIAL GROUP senses, which they clearly possess (cf. *The whole laboratory celebrated Christmas*). This suggests that our approach can help in Word Sense Induction and thesaurus construction.

In the second case, meta alternations perform better than expected: They have a low κ, but a high AP. These include grs-psy, pro-sta, psy-sta, hum-prt and grp-psy. These meta alternations involve fairly abstract meta senses such as PSYCHOLOGICAL FEATURE and STATE.[9] Table 6 lists a sample of targets for the five meta alternations involved. The targets are clearly similar to each other on the level of their meta senses. However, they can occur in very different semantic contexts. Thus, here it is the underlying model (the gramlex space) that can explain the lower than average coherence. It is striking that CAM can account for abstract words and meta alternations between these, given that it uses first-order co-occurrence information only.

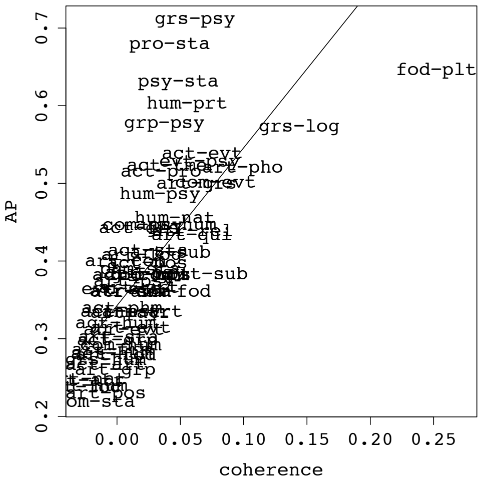

Figure 2: Average Precision and Coherence (κ) for each meta alternation. Correlation: $r = 0.743$ ($p < 0.001$)

6 Related work

As noted in Section 1, there is little work in empirical computational semantics on explicitly modeling sense alternations, although the notions that we have formalized here affect several tasks across NLP subfields.

Most work on regular sense alternations has focused on regular polysemy. A pioneering study is Buitelaar (1998), who accounts for regular polysemy through the CoreLex resource (cf. Section 3). A similar effort is carried out by Tomuro (2001), but he represents regular polysemy at the level of senses. Recently, Utt and Padó (2011) explore the differences between between idiosyncratic and regular polysemy patterns building on CoreLex. Lapata (2000) focuses

[9] An exception is hum-prt. It has a low coherence because many WordNet lemmas with a PART sense are body parts.

on the default meaning arising from word combinations, as opposed to the polysemy of single words as in this study.

Meta alternations other than regular polysemy, such as metonymy, play a crucial role in Information Extraction. For instance, the meta alternation SOCIAL GROUP-GEOGRAPHICAL LOCATION corresponds to an ambiguity between the LOCATION-ORGANIZATION Named Entity classes which is known to be a hard problem in Named Entity Recognition and Classification (Markert and Nissim, 2009). Metaphorical meta alternations have also received attention recently (Turney et al., 2011)

On a structural level, the prediction of meta alternations shows a clear correspondence to analogy prediction as approached in Turney (2006) (*carpenter:wood* is analogous to *mason:stone*, but not to *photograph:camera*). The framework defined in Section 2 conceptualizes our task in a way parallel to that of analogical reasoning, modeling not "first-order" semantic similarity, but "second-order" semantic relations. However, the two tasks cannot be approached with the same methods, as Turney's model relies on contexts linking two nouns in corpus sentences (*what does A do to B*?). In contrast, we are interested in relations *within* words, namely between word senses. We cannot expect two different senses of the same noun to co-occur in the same sentence, as this is discouraged for pragmatic reasons (Gale et al., 1992).

A concept analogous to our notion of meta sense (i.e., senses beyond single words) has been used in previous work on class-based WSD (Yarowsky, 1992; Curran, 2005; Izquierdo et al., 2009), and indeed, the CAM might be used for class-based WSD as well. However, our emphasis lies rather on modeling polysemy across words (meta alternations), something that is absent in WSD, class-based or not. The only exception, to our knowledge, is Ando (2006), who pools the labeled examples for all words from a dataset for learning, implicitly exploiting regularities in sense alternations.

Meta senses also bear a close resemblance to the notion of semantic class as used in lexical acquisition (Hindle, 1990; Merlo and Stevenson, 2001; Schulte im Walde, 2006; Joanis et al., 2008). However, in most of this research polysemy is ignored. A few exceptions use soft clustering for multiple assignment of verbs to semantic classes (Pereira et al., 1993; Rooth et al., 1999; Korhonen et al., 2003), and Boleda et al. (to appear) explicitly model regular polysemy for adjectives.

7 Conclusions and Future Work

We have argued that modeling regular polysemy and other analogical processes will help improve current models of word meaning in empirical computational semantics. We have presented a formal framework to represent and operate with regular sense alternations, as well as a first simple instantiation of the framework. We have conducted an evaluation of different implementations of this model in the new task of determining whether words match a given sense alternation. All models significantly outperform the baselines when considered as a whole, and the best implementation outperforms the baselines for 73.3% of the tested alternations.

We have two next steps in mind. The first is to become independent of WordNet by unsupervised induction of (meta) senses and alternations from the data. This will allow for models that, unlike CAM, can go beyond "disemous" words. Other improvements on the model and evaluation will be to develop more informed baselines that capture semantic shifts, as well as to test alternate weighting schemes for the co-occurrence vectors (e.g. PMI) and to use larger corpora than the BNC.

The second step is to go beyond the limited in-vitro evaluation we have presented here by integrating alternation prediction into larger NLP tasks. Knowledge about alternations can play an important role in counteracting sparseness in many tasks that involve semantic compatibility, e.g., testing the applicability of lexical inference rules (Szpektor et al., 2008).

Acknowledgements

This research is partially funded by the Spanish Ministry of Science and Innovation (FFI2010-15006, TIN2009-14715-C04-04), the AGAUR (2010 BP-A00070), the German Research Foundation (SFB 732), and the EU (PASCAL2; FP7-ICT-216886). It is largely inspired on a course by Ann Copestake at U. Pompeu Fabra (2008). We thank Marco Baroni, Katrin Erk, and the reviewers of this and four other conferences for valuable feedback.

References

Rie Kubota Ando. 2006. Applying alternating structure optimization to word sense disambiguation. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 77–84, New York City, NY.

Iurii Derenikovich Apresjan. 1974. Regular polysemy. *Linguistics*, 142:5–32.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. to appear. Modeling regular polysemy: A study of the semantic classification of Catalan adjectives. *Computational Linguistics*.

Paul Buitelaar. 1998. CoreLex: An ontology of systematic polysemous classes. In *Proceedings of Formal Ontologies in Information Systems*, pages 221–235, Amsterdam, The Netherlands.

Gavin Burnage and Dominic Dunlop. 1992. Encoding the British National Corpus. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: Design, Analysis and Exploitation, Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*. Rodopi, Amsterdam.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with ccg and log-l inear models. *Computational Linguistics*, 33(4).

Ann Copestake and Ted Briscoe. 1995. Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 12(1):15–67.

James Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 26–33, Ann Arbor, Michigan.

Bradley Efron and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT, London.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the 1992 ARPA Human Language Technologies Workshop*, pages 233–237, Harriman, NY.

Dedre Gentner, Brian F. Bowdle, Phillip Wolff, and Consuelo Boronat. 2001. Metaphor is like analogy. In D. Gentner, K. J. Holyoak, and B. N. Kokinov, editors, *The analogical mind: Perspectives from Cognitive Science*, pages 199–253. MIT Press, Cambridge, MA.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Meeting of the Association for Computational Linguistics*, pages 268–275.

Rubén Izquierdo, Armando Suárez, and German Rigau. 2009. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 389–397, Athens, Greece.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Anna Korhonen, Yuval Krymolowski, and Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Mirella Lapata. 2000. *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Ph.D. thesis, University of Edinburgh.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting on Association for Computational Linguistics*, pages 25–32, College Park, MA.

Will Lowe. 2001. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pages 576–581, Edinburgh, UK.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 1st edition.

Katja Markert and Malvina Nissim. 2009. Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Using automatically acquired predominant senses for word sense disambiguation. In *Proceedings of the ACL SENSEVAL-3 workshop*, pages 151–154.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

Gregory L. Murphy. 2002. *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086, July.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41:10:1–10:69, February.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002*, pages 613–619, Edmonton.

Fernando C. N. Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)*, pages 109–117.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 683–691, Columbus, Ohio.

Noriko Tomuro. 2001. Tree-cut and a lexicon based on systematic polysemy. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32:379–416.

Jason Utt and Sebastian Padó. 2011. Ontology-based distinction between polysemy and homonymy. In *Proceedings of the 9th International Conference on Computational Semantics*, Oxford, UK.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 454–460, Stroudsburg, PA, USA. Association for Computational Linguistics.