Class-based Word Sense Induction for dot-type nominals

Lauren Romeo
Universitat Pompeu Fabra
Roc Boronat, 138
Barcelona (Spain)
lauren.romeo@upf.edu

Héctor Martínez Alonso
University of Copenhagen
Njalsgade, 140
Copenhagen (Denmark)
alonso@hum.ku.dk

Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138
Barcelona (Spain)
nuria.bel@upf.edu

Abstract

This paper describes an effort to capture the sense alternation of dot-type nominals using Word Sense Induction (WSI). We propose dot-type nominals generate more semantically consistent groupings when clustered into more than two clusters, accounting for literal, metonymic and underspecified senses. Using a class-based approach, we replace individual lemmas with a placeholder representing the entire dot type, which also compensates for data sparsity. Although the distributional evidence does not motivate an individual cluster for each sense, we discuss how our results empirically support theoretical proposals regarding dot types.

1 Introduction

In this article, we propose a Word Sense Induction (WSI) task to capture the sense alternation of English dot types, as found in context. *Dot type* is the Generative Lexicon (GL) term to account for a noun that can denote at least two senses as a complex semantic class (Pustejovsky, 1995). Consider the noun *England* in the following example from the American National Corpus (ANC) (Ide and Macleod, 2001) as an illustration.

(1) (a) Manuel died in exile in 1932 in *England*.
(b) *England* was being kept busy with other concerns.
(c) *England* is conservative and rainy.

In this example, (1a) shows the *literal* sense of England as a location, while (1b) demonstrates the *metonymic* sense of England as an organization. Dot types also allow for both senses to be simultaneously active in a predicate, as in example (1c). All proper names representative of geopolitical entities, for instance, demonstrate this type of class-wide sense alternation, which is defined as *regular polysemy* (Apresjan, 1974).

Copestake (2013) emphasizes the relevance of distributional evidence in tasks regarding phenomena characteristic to regular polysemy, such as underspecification, because it incorporates frequency effects and is theory-neutral, requiring only that examples cluster in a way that mirrors their senses.

Thus far, underspecification in dot types has been formalized in the linguistic theory of lexical semantics, but has not been explicitly studied using WSI. Kilgariff (1997) claims that word senses should be *"construed as abstractions over clusters of word usages"*. Following this claim, our strategy employs WSI, which aims to automatically induce senses of words by clustering patterns found in a corpus (Lau et al., 2012; Jurgens, 2012). In this way, we hypothesize that dot-type nominals will generate semantically more consistent (i.e. more homogeneous, cf. Section 5) groupings if clustered into more than two induced senses.

This paper is organized as follows: we discuss related work (Section 2); elaborate upon our use of WSI and methodology employed (Section 3 and Section 4), as well as present results obtained; we discuss our results (Section 5) and conclude with final observations and future work (Sections 6 and 7).

2 Related Work

Natural Language Processing (NLP) tasks that exploit distributional information are based on the Distributional Hypothesis (Harris, 1954). However, Pustejovsky and Ježek (2008) claim that only using distributional data cannot explain the variation of linguistic meaning in language, while Markert and Nissim (2009) refer to the challenges of dealing with regular polysemy as the different senses of polysemous words present obstacles due to varied use in context. Along this line, the empirical work of Boleda et al. (2012) showed that the skewed sense distribution of many words makes it difficult to distinguish evidence of a class from noise, presenting a challenge to model the relations between senses. When their machine-learning experiments reached the upper bound set by the inter-encoder agreement in their gold standard, they concluded that in order to improve the modelling of polysemy there is a need to shift from a type to a token-based (word-in-context) model (Schütze, 1998; Erk and Padó, 2008). Hence, we employ a token-based model in our experiments.

In our approach, we propose an unsupervised task using WSI to capture the sense alternation of dot types, using distributional evidence from corpus data. Our results will be noisier than supervised approaches, such as those of Markert and Nissim (2009), Nissim and Markert (2005) and Nastase et al. (2012), but we make use of a much larger amount of data and thus should suffer from less sparsity. The related experiment by Rumshisky et al. (2007) uses verbal arguments as features, while we use only a five-word context window.

2.1 Word Sense Induction

As stated above, our main goal is to use WSI to capture the sense alternation of dot types in context. WSI methods, based on the distributional information available in corpus data, employ unsupervised means to induce senses using contexts of indicated target words without relying on handcrafted resources (Manandhar et al., 2010).

Distributional Semantic Models (DSM) provide the groundwork for WSI. A DSM, also known as a Word Space Model (Turney and Pantel, 2010), attempts to describe the meaning of words by characterizing their usage over distributional patterns, i.e. their context. Each word is represented by a numeric vector positioned in a space where vectors for words that appear in similar contexts are closer to each other. Sense induction is achieved by building a DSM over a large corpus and clustering the contexts into induced senses.

In recent years, WSI has been used with success for different tasks such as: novel sense detection (Lau et al., 2012), community detection (Jurgens, 2011) and graded sense disambiguation (Jurgens, 2012), among others. Jurgens (2011) previously employed WSI to discover overlaps in the distributional behavior of words in order to identify multiple senses with success. However, that work was not inclusive to any specific phenomenon of polysemy. Our objective is to cluster dot-type nominals according to their distributional evidence in context, using WSI to characterize the behavior of these nouns.

3 Method

We use WSI to computationally assess the predicational behavior of dot types. To do this, we employ a WSI system to induce senses from a large corpus (in our case UkWaC cf. Section 3.2). We then cluster dot-type nominals into the different induced k-solutions and evaluate the WSI model using a dot-type sense-annotated corpus to measure how well the induced senses map to human-annotated data.

3.1 Data

The dot-type sense-annotated corpus (Martínez Alonso et al., 2013) provides examples for each of the following dot types:

1. Animal/Meat (ANIMEAT): *The chicken ran away* vs. *the chicken was delicious.*
2. Artifact/Information (ARTINFO): *The book fell* vs. *the book was boring.*
3. Container/Content (CONTCONT): *The box was red* vs. *I ate the whole box.*
4. Location/Organization (LOCORG): *England is far* vs. *England starts a tax reform.*
5. Process/Result (PROCRES): *The building took months to finish* vs. *the building is sturdy.*

To evaluate our clustering, we made use of the aforementioned sense-annotated corpus as a gold standard. The corpus provides senses that have been obtained by majority voting with a theory-compliant back-off strategy (see Martínez et al., 2013 for a detailed description). Each section of the sense-annotated corpus[1] is a block of 500 sentences with one dot-type headword the annotators had to disambiguate. The authors do not make a distinction between sense alternations that are based on physical contiguity (CONTCONT) from temporal contiguity (PROCRES). We use their data as provided.

The gold standard includes nouns annotated as literal, metonymic or underspecified. Each dataset

[1] We obtained the data from MetaShare at http://metashare.cst.dk/repository/search/?q=regular+polysemy

Dot type	$\overline{A_o}$	α
ANIMEAT	0.86	0.69
ARTINFO	0.48	0.12
CONTCONT	0.65	0.31
LOCORG	0.72	0.46
PROCRES	0.50	0.10

Table 1: Averaged observed agreement ($\overline{A_o}$) and Krippendorf's alpha (α)

has a different average observed agreement and Krippendorf's α coefficient (cf. Poesio and Artstein, 2008), as shown in Table 1.

The variation in agreement for each dataset was strong, which is a sign of the difficulty of each annotation task. For instance, LOCORG is easier to annotate than ARTINFO, which is reflected in its higher agreement. Another relevant characteristic of the gold standard is that there is also an imbalance of frequency between the annotated senses of each dot type. For instance, it resulted that ANIMEAT was annotated with more literal readings and PROCRES was annotated with more metonymic readings. Figure 1 provides the distribution of senses between each dot type studied in this article.

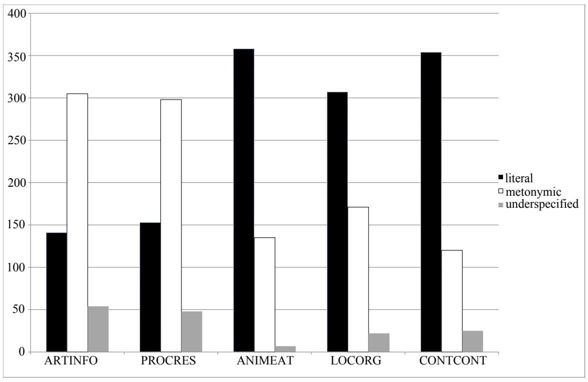

Figure 1: Distribution of senses between classes

3.2 Preprocessing

For our experiments we used the UkWaC corpus (Baroni et al., 2009) to fit our WSI models. After lemmatizing, lowercasing and removing all punctuation from the corpus, we extracted a random sample of 60 million words (2.8 million sentences) where each sentence was at least five tokens long. We did not remove stop words from the corpus as we expect the interaction between stop words (e.g. articles, prepositions, etc.) and dot-type nominals to represent strong distinguishing features between different interpretations of a dot type, along the lines of Rumshisky et al. (2007).

In our experiments, we assume that words of the same class behave similarly. Thus, our intent is to induce the same senses for all the words of a given semantic class, making our approach class-based.

To group the occurrences of all words of a given dot type, we replaced their occurrences with a placeholder lemma that represents the entire dot type (*animeatdot, artinfodot, contcontdot, locorgdot, procresdot*). For instance, the lemmatized examples (2a) and (2b) with the words *paris* and *london* become the sentences in the examples (2c) and (2d).

(2) (a) whilst i be in **paris** in august i decide to visit the catacomb
(b) you can get to both **london** station on the **london** underground
(c) whilst i be in **locorgdot** in august i decide to visit the catacomb
(d) you can get to both **locorgdot** station on the **locorgdot** underground

Replacing individual lemmas by a placeholder for the overall class yields results similar to those obtained by building prototype distributional vectors for a set of words once the DSM has been calculated (cf. Turney and Pantel (2010) for more on prototype vectors of a semantic class). Our take, however, is a preprocessing of the corpus to assure we infer senses directly for the placeholder lemmas. In this way, we avoid having to reconstruct overall class-wise senses from the inferred senses for each individual lemma.

Regular polysemy is a class-wide phenomenon (cf. Section 1), hence we expect that all lemmas in a dot type will predicate their senses in a similar manner—in similar contexts, e.g. headed or followed by the same prepositions. Thus, the placeholders represent the entire dot type as well as provide the added benefit of circumventing the effects of data sparseness, especially for evaluation purposes. For instance, in our data there are some lemmas (eg. in ANIMEAT: *anchovy, yak, crayfish*) that only appear once in the gold standard, limiting evaluation power. The placeholder reduces the impact this may have on evaluation by considering each individual lemma as a member of the entire dot type that its placeholder represents.

This replacement method is not exhaustive because we strictly replace the words from the test dataset by their dot-type placeholder and, for instance, plenty of country and city names are not replaced by *locorgdot* as they were not considered target nouns in the annotation task.

3.3 Applying WSI

Our WSI models were built using the Random Indexing Word Sense Induction module in the S-Spaces package for DSMs (Jurgens and Stevens, 2010) employing the UkWaC corpus, as described in Section 3.2. Random Indexing (RI) is a fast method to calculate DSMs, which has proven to be as reliable as other word-to-word DSMs, like COALS (Rohde et al., 2009). In DSMs, words are represented by numeric vectors calculated from the occurrence of words in a n-word window around a target word. The similarity between words is measured by means of the cosine of the vectors that represent them.

We induced the senses for the placeholder dot-type lemmas (*locorgdot*, *animeatdot*, and so on), using the following k values to see how the senses are clustered when considering a coarse (k=2; literal and metonymic), a medium (k=3; literal, metonymic, underspecified) and a finer-grained amount of induced senses (k=6), along the lines of Markert and Nissim (2009).

In WSI, instead of generating one vector for each word, each word is assigned k vectors, one for each induced sense. These induced vectors are obtained by clustering the occurrences of a selected word into k senses. The features used to cluster the contexts into senses were the words found in a window of five, both to the left and the right of the target word. For each of the three values of k, we fit a model using K-means clustering and a model using Spectral Clustering (Cheng et al., 2006), for a total of 6 models. The output of the system is a DSM where each vector is one of the k-induced senses for the placeholder dot-type lemmas.

3.4 Assigning word senses

The S-Spaces API permits the calculation of a vector in a DSM for a new, unobserved example. For each sentence in the test data, we isolated the placeholder to disambiguate and we calculated the representation of the sentence within the corresponding WSI model using the specified 5-word context window.

Once the vector for the sentence was obtained, we assigned the sentence to the induced sense representing the highest cosine similarity for each model (cf. Table 2 in Section 4 for evaluation).

4 Results

To determine the success of our task for each class, sense representation and k value, we consider the information-theoretic measures of *homogeneity*, *completeness* and *V-measure* (Rosenberg and Hirschberg, 2007). These three measures compare the output of the clustering with a gold standard (as described in Section 3.1) and provide a score that can be interpreted in a manner similar to precision, recall and F1, respectively.

Homogeneity determines to which extent each cluster only contains members of a single class, while completeness determines if all members of a given class are assigned to the same cluster. Both the homogeneity and completeness scores are bounded by 0.0 and 1.0, with 1.0 corresponding to the most homogeneous or complete solution, and can be interpreted in a manner similar to precision and recall.

V-measure is the harmonic mean of homogeneity and completeness, used to evaluate the agreement of two independent assignments on the same dataset. Values close to zero indicate two label assignments that are largely inconsistent, while values close to one indicate consistency. Much like F1, the V-score indicates the best trade-off between homogeneity and completeness.

	DATASET	**HOM**	**COM**	**V-ME**
k=2	ANIMEAT	0.0031	0.0030	0.0030
	ARTINFO*	**0.0097**	**0.0128**	**0.0110**
	CONTCONT*	0.0067	**0.0075**	0.0071
	LOCORG*	0.0013	0.0016	0.0015
	PROCRES	0.0005	0.0007	0.0006
k=3	ANIMEAT	0.0055	0.0033	0.0041
	ARTINFO*	0.0214	0.0191	0.0201
	CONTCONT*	0.0291	0.0197	**0.0235**
	LOCORG*	**0.1070**	**0.0788**	**0.0908**
	PROCRES*	0.0051	0.0044	0.0047
k=6	ANIMEAT*	0.0379	0.0139	0.0204
	ARTINFO*	0.0253	0.0140	0.0180
	CONTCONT*	**0.1008**	0.0442	**0.0615**
	LOCORG*	**0.1096**	**0.0540**	**0.0724**
	PROCRES*	0.0166	0.0085	0.0112

Table 2: Results of clustering solutions for each class in terms of homogeneity **(HOM)**, completeness **(COM)** and V-measure **(V-ME)**

Table 2 presents the results for each clustering solution (k=2, k=3 and k=6) using K-means clustering. The highest values are shown in bold. It is to be expected that the higher-agreement datasets provide higher homogeneity results because their annotations are more consistent. However, we can see that the performance does not necessarily correlate with agreement as ARTINFO is the dataset that fares best in the k=2 solution, yet it has a very low alpha (α=0.12). In this way, we can say that the homogeneity score for low-agreement datasets will be lower because low-agreement annotations are less reliable due to their lower internal consistency.

In addition, performance (measured in V-measure) improves as k increases. For instance, CONTCONT has the 2nd highest V-measure in the k=3 solution and in the k=6 solution. LOCORG yielded the 4th highest V-measure in k=2 and the highest V-measure in both the k=3 and the k=6 solutions.

We compare our system against a random baseline. This is because the customary one-in-all and all-in-one baselines are not useful in our scenario as they are meant to evaluate adaptative clustering and we use fixed values of K. We do not report the baseline scores because they are not informative. However, we mark the datasets that surpassed those scores with a star (*) in Table 2.

Although our system is unable to beat the random baseline for PROCRES in k=2 and ANIMEAT for k=2 and k=3, we do beat the baseline for each dot type in k=6.

The low performance in ANIMEAT is due to the lower proportion of underspecified senses in the dataset (cf. Figure 1). We attribute the low performance of PROCRES to the complexity of the sense distinction of this dot type. Thereby, we doubt the validity of this particular dataset for WSI.

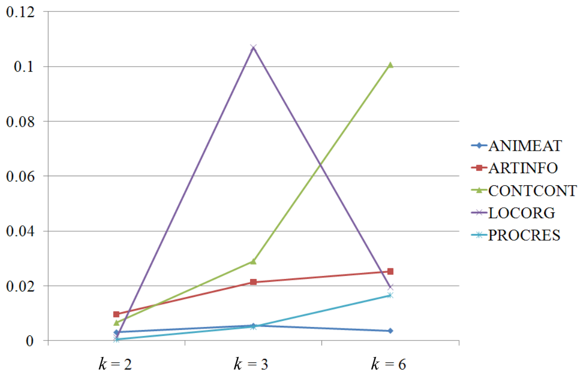

Figure 2: Homogeneity scores for each clustering solution

Figure 2 demonstrates the difference of homogeneity between the clusters, depending on the number of induced senses (k-value). LOCORG and ANIMEAT, on one hand, demonstrate a higher homogeneity score in k=3 while they demonstrate a lower homogeneity score for k=6. CONTCONT, ARTINFO, PROCRES, on the other hand, gain homogeneity with the increase of k.

5 Discussion

The main objective of this experiment is to capture the sense alternation of dot types by computational means. We hypothesize that dot types will generate semantically more consistent groupings if clustered into more than two clusters. To test this, we employ a WSI system to induce the senses and subsequently cluster dot-type nominals into three different k solutions (k=2, k=3, k=6), as detailed in 3.3.

5.1 Inducing two senses

The k=2 solution attempts to mirror a literal vs. metonymic partition between the senses of each dot type. The classes ANIMEAT, CONTCONT and LOCORG are composed of more literal senses while the other two are mostly metonymic (cf. Figure 1). Although there is an a priori difference in the proportion of literal, metonymic and underspecified senses for each class, we assume the UkWaC and test data to have similar distributions of literal and metonymic senses for each dot type. This assumption is congruent with Rumshisky et al. (2007), who claim an asymmetry in the way dot types are used in general.

Overall, the clusters produced in k=2, on one hand, are representative of the asymmetry of the gold standard, i.e. the classes that contain more literal senses, according to our gold standard, yield clusters composed of a higher ratio of literal senses. On the other hand, the underspecified senses tend to spread between both clusters for each class. In this way, the underspecified sense does not represent a homogeneous group, rather it clusters with both the literal and metonymic senses, thereby exhibiting properties of each of the two induced senses.

We observed, for instance, the underspecified senses of ARTINFO occurred often with an “of” PP-phrase, a strong feature for the clustering of examples into a metonymy-dominated sense cluster while the underspecified examples that were objects of verbs such as *keep* or *see* were clustered alongside the literal examples. In this way and along the lines of Pustejovsky and Ježek (2008), we can concur that these verbs tend to trigger a literal (artifactual) reading as they typically describe actions that require some sort of physical entity.

We next increase the K to k=3, a solution that also considers the underspecified sense.

5.2 Inducing three senses

The goal of the k=3 clustering solutions is to cluster each of the three proposed senses of the dot type (literal, metonymic and underspecified) into clusters representative of their respective senses.

The middle row in Table 2 presents the results obtained in the k=3 solution. Our expectation for this solution would have had each gold-standard annotated sense assigned to its corresponding induced sense cluster (literal, metonymic or underspecified). However, we noticed a tendency for the underspecified sense to cluster with the induced sense that contains a higher ratio of the most frequent annotated sense of a given class, either literal or metonymic. Despite the fact that the distributional information for the underspecified sense was not strong enough to spawn a separate cluster, it demonstrates behavior characteristic of the more frequent sense for each dot type, as indicated by the gold standard (cf. Figure 1).

The k=3 solution for the dot types ANIMEAT and LOCORG separates the literal and the metonymic senses, yet the underspecified senses are distributed between all three clusters. In this case, the more frequent sense of the gold standard is split between two clusters, while the remaining cluster is composed of the less frequent sense. The underspecified sense is spread among all three clusters, as illustrated in the confusion matrices provided in Table 3.

	ANIMEAT			LOCORG		
	L	M	U	L	M	U
c=0	110	51	3	62	69	8
c=1	127	43	1	151	17	5
c=2	121	41	3	94	85	9

Table 3: k=3 solutions for ANIMEAT and LOCORG dot types

In Table 4, we observed that the articles *the* and *a* were the most frequent components of the contexts that contributed to the clustering of clusters c=1 and c=2, respectively, for ANIMEAT. On one hand, the importance of the article as a feature reflects that the mass/count distinction is a key component in the sense alternation of some instances of regular polysemy (such as ANIMEAT). In this way, these very formalized constructs that are required for a certain interpretation can help to more easily partition the clusters as they represent grammatical criteria for interpretation. On the other hand, we can also see the importance of a given token in context in the case of LOCORG. For LOCORG, in c=1 and c=2, the most frequent components that contribute to each cluster are prepositions (*in* and *to*, respectively).

	ANIMEAT	LOCORG
c=0	*and*, *animeatdot*, *of*, *a*, *for*,*with*, *the*, *fish*, *in*, *to*	*the*, *of*, *and*, *to*, *a*, *in*, *that*, *time*, *it*, *for*
c=1	*the*, *of*, *and*, *in*, *a*, *to*, *is*, *that*, *animeatdot*, *with*	*in*, *the*, *and*, *to*, *a*, *of*, *that*, *is*, *it*, *for*
c=2	*a*, *of*, *to*, *in*, *that*, *or*, *is*, *with*, *for*, *from*	*to*, *and*, *from*, *a*, *locorgdot*, *the*, *that*, *with*, *for*, *is*

Table 4: Top 10 most frequent words per c used in k=3 for ANIMEAT and LOCORG dot types

The very frequent preposition *in* seems to favor the literal (*location*) reading for LOCORG that appears in c=1. In c=2, the most important preposition is *to*, which indicates a directionality that can be both topological or more abstract, giving to the introduced noun the role of experiencer or beneficiary in the predicate, for instance. However, this preposition does not necessarily coerce a metonymic or a literal sense, which becomes apparent in the balanced composition of the senses of LOCORG in c=2.

The placeholders also appear as important features for their respective dot type among all the grammatical words. We observed that other animals are mentioned when predicating the ANIMEAT dot type (see Table 4). The noun *fish* was not replaced by its placeholder as it does not appear in the gold standard data but is one of the few nouns in the top 10 words for each cluster. We comment upon the effect of our use of a limited selection of lemmas in this task in Section 7.

Overall, the distributional evidence used in the k=3 solution is again not strong enough to motivate an individual cluster for each sense, indicating the underspecified senses may not be as lexically homogeneous as the other two. This is because they have properties of both senses of a given dot type, supporting the assumption that the underspecified sense is formed by the union of both the literal and metonymic senses (Pustejovsky, 1995). However, under the assumption that more fine-grained patterns may indicate underspecified reading, we attempted a k=6 solution to differentiate between senses with a larger K.

5.3 Inducing six senses

The k=6 solution was proposed to uncover fine-grained sense distinctions between a given dot type (Markert and Nissim, 2009). We observed, namely in CONTCONT, ANIMEAT and PROCRES, that the resulting clusters demonstrate a higher V-measure than their k=2 and k=3 counterparts, but this is a consequence of a higher homogeneity expected from an increased k-value. On one hand, the less homogeneous clusters in k=3 are more prone to be split into at least two smaller yet more homogeneous clusters in k=6. On the other hand, the more homogeneous clusters in k=3 were mostly preserved in k=6, as the senses that pertain to it remained identifiable in its own separate cluster. This demonstrates that, although disperse, the resulting clusters contain stable elements that are representative of a given sense.

The k=6 solution is thus a further refinement of k=3 into more fine-grained induced senses. The results for k=6 still reflect the challenges of the task and the variation of the sense composition of dot-type nominals, i.e. they occur predominantly in one sense and the distributions of their underspecified senses largely overlap with the distribution of the literal and metonymic senses.

6 Conclusions

In this work, our objective was to use WSI to capture the sense alternation of dot types. Although our system surpassed the random baseline for all dot types in k=6, the V-measure of the induced-sense clustering solutions demonstrates that our method was not able to isolate the literal, metonymic and underspecified senses. Our results do not imply an absolute distinction between the senses of a dot type.

The skewedness in sense distributions of the dot types in the gold standard (cf. Figure 1) has an impact on the quality of our results. This can be attributed to a preference of a dot type to be selected for more often as one sense over the other in a given context, along the lines of Rumshisky et al. (2007).

The lower-agreement datasets (cf. Table 1; CONTCONT, PROCRES) increase in homogeneity with the increase of K (see Table 2), suggesting that more difficult-to-annotate dot types have more variation and thus cluster better in a higher K.

The differences between the contexts of the senses were still not strong enough to motivate separate clusters for each individual sense. This is in line with Markert and Nissim (2009) and Boleda et al. (2012) which refer to the difficulty of dealing with different forms of regular polysemy as a factor that limits conclusion power. It is also in line with Pustejovksy and Ježek (2008), as our analysis provided distributional evidence considering only 5-word window contexts, which do not reflect modulations that a given dot type may undergo due to its occurrence in context. We leave the refinement of features for future work (see Section 7).

7 Future Work

In many cases the clustering solutions appear to be governed by a particular syntactic or lexical context (i.e. a dependent PP in the case of the metonymic-dominated cluster of ARTINFO), denoting its resulting sense through a specific context. Moreover, our DSM only calculated relations between lemmas. However, we are aware, for instance, that the plural number is an informative feature for the count/mass alternation (Gillon, 1992), which is parallel to many instances of regular polysemy (Copestake, 2013).

As we use 5-word contexts to induce and subsequently cluster our senses, we do not capture all the contextually complex phrases or gating predicates, coordinated co-predications, and vague contexts that can cause underspecified predications. However, our results depend not only on an accurate induction of the senses in context, but also on the reliability of the test set (see Table 1).

We also consider that we now have a baseline which provides information with regard to the sense relations of a given dot type, as per our analysis based on the results of our WSI task. Thereby, we can use a DSM for a WSI that takes into account syntactic role of each token to compare results.

Finally, the placeholder lemmas replace all the lemmas in the gold standard, as indicated in Section 3.2. The selection of lemmas that we replace restricts the class-based WSI because of its small sample size. We should expand these lists with more lemmas, so the distribution of the semantic class can be less biased by the choice of lemmas.

Acknowledgements

This work was funded with the support of the SUR of the DEC of the Generalitat de Catalunya and the European Social Fund, by the European Commission's 7th Framework Program under grant agreement 238405 (CLARA) and by SKATER TIN2012-38584-C06-05.

References

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

G. Boleda, S. Schülte im Walde, and T. Badia. 2012. Modeling regular polysemy: A study of the semantic classification of catalan adjectives. *Computational Linguistics*, 38(3):575–616.

D. Cheng, R. Kannan, S. Vempala, and G. Wang. 2006. A divide-and-merge methodology for clustering. *ACM Transactions on Database Systems (TODS)*, 31(4):1499–1525.

A. Copestake. 2013. Can distributional approached improve on goold old-fashioned lexical semantics? In *IWCS Workshop Towards a Formal Distributional Semantics*.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics.

B.S. Gillon. 1992. Towards a common semantics for english count and mass nouns. *Linguistics and Philosophy*, 15:597–639.

Z. Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

N. Ide and C. Macleod. 2001. The american national corpus: A standardized resource of american english. In *Corpus Linguistics*, pages 274–280.

D. Jurgens and K. Stevens. 2010. Measuring the impact of sense similarity on word sense induction. In *First workshop on Unsupervised Learning in NLP (EMNLP 2011)*.

D. Jurgens. 2011. Word sense induction by community detection. In *6th ACL Workshop on Graph-based Methods for Natural Language Processing (Text-Graphs 6)*.

D. Jurgens. 2012. An evaluation of graded sense disambiguation using word sense induction. In **SEM First Joint Conference on Lexical and Computational Semantics*.

A. Kilgariff. 1997. I dont believe in word senses. *Computers and the Humanities*, 31:91–113.

J.H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. 2012. Word sense induction for novel sense detection. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

S. Manandhar, I.P. Klapaftis, D. Dligach, and S. Pradhan. 2010. Task 14: Word sense induction and disambiguation. In *15th International Workshop on Semantic Evaluation (ACL)*, pages 63–68.

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

H. Martínez Alonso, B. Sandford Pedersen, and N. Bel. 2013. Annotation of regular polysemy and underspecification. In *51st Meeting of the Associatation for Computation Linguistics (ACL 2013)*.

V. Nastase, A. Judea, K. Markert, and M. Strube. 2012. Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 183–193. Association for Computational Linguistics.

M. Nissim and K. Markert. 2005. Learning to buy a renault and talk to bmw: A supervised approach to conventional metonymy. In *International Workshop on Computational Semantics (IWCS2005)*.

J. Pustejovsky and E. Ježek. 2008. Semantic coercion in language: Beyond distributional analysis. *Italian Journal of Linguistics*.

J. Pustejovsky. 1995. *The Generative Lexicon*. Oxford University Press, Oxford.

D. Rohde, L. Gonnerman, and D. Plaut. 2009. An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*.

A. Rosenberg and J. Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2007)*.

A. Rumshisky, V. Grinberg, and J. Pustejovsky. 2007. Detecting selectional behavior of complex types. In *4th International Workshop on Generative Approaches to the Lexicon*.

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

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