A Word-Embedding-based Sense Index for Regular Polysemy Representation

Marco Del Tredici
Universitat Pompeu Fabra
Roc Boronat, 138
Barcelona, Spain
marco.deltredici@upf.edu

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

Abstract

We present a method for the detection and representation of polysemous nouns, a phenomenon that has received little attention in NLP. The method is based on the exploitation of the semantic information preserved in Word Embeddings. We first prove that polysemous nouns instantiating a particular sense alternation form a separate class when clustering nouns in a lexicon. Such a class, however, does not include those polysemes in which a sense is strongly predominant. We address this problem and present a sense index that, for a given pair of lexical classes, defines the degree of membership of a noun to each class: polysemy is hence implicitly represented as an intermediate value on the continuum between two classes. We finally show that by exploiting the information provided by the sense index it is possible to accurately detect polysemous nouns in the dataset.

1 Introduction

A major issue in lexical semantics is regular polysemy (also known as *systematic* or *logical* polysemy), the phenomenon whereby words belonging to a semantic class can predictably act as members of another class (Pustejovsky, 1991; Martínez Alonso et al., 2013). For example, the word *chicken* can be considered a member of the class ANIMAL but also of FOOD, thus defining its senses in terms of lexical semantic classes. For some polysemous nouns one sense can be much more frequent than the other, thus causing asymmetry in sense predominance; this is the case of *turkey*, in which the food sense is clearly more frequent than the animal one (Copestake and Briscoe, 1995).

Given its pervasiveness in natural language, regular polysemy has been extensively investigated in lexical semantics (Apresjan, 1974; Nunberg, 1992). However, only few works attempted to computationally model this phenomenon (Copestake, 2013; Boleda et al., 2012b). The vast majority of applications treat regular polysemy like other phenomena of lexical ambiguity, such as homography, not considering the relevant theoretical differences between those phenomena, for example that regular polysemy is predictable, while homography is not (Utt and Padó, 2011). Information on regular polysemy would be valuable for a task like Word Sense Disambiguation, since it would reduce the number of possible options when choosing the right sense of a word. More generally, every lexical resource would benefit from the capability to cope with the shifts of meaning produced by regular polysemy, and this, in turn, would lead to improvements in several NLP applications as such machine translation, textual entailment or text analytics.

In this paper we present a method for polysemy detection and representation based on Word Embeddings (WE) (Mikolov et al., 2013a). WE are low dimensional, dense and real-valued vectors which preserve word syntactic and semantic information in a Vector Space Model (VSM). WE have recently been proved to be efficient in several NLP tasks, such as detection of relational similarity (Mikolov et al., 2013d), word similarity tasks (Mikolov et al., 2013a) and automatic building of bilingual lexica (Mikolov et al., 2013b), in which this word repre-

Proceedings of NAACL-HLT 2015, pages 70–78,
Denver, Colorado, May 31 – June 5, 2015.

sentations outperformed others with state-of-the-art methods (Baroni et al., 2014).

However, at the best of our knowledge, this is the first work in which WE are used to represent and account for regular polysemy. Our work departs from the assumption that lexical classes related by regular polysemy are limited and known and that since the class-related senses of a polysemous nouns can be considered as modulations of meaning of a single lemma (Copestake and Briscoe, 1995), they are to be represented by a single vector. As a first step, we will prove through a clustering task that nouns instantiating a particular sense alternation (e.g. animal/food) group together and separately from non-polysemous nouns, forming a distinct cluster. Such a cluster, however, does not include polysemous nouns with a strong sense asymmetry. Therefore, obtained clustering information is exploited in order to assign each noun a *sense index*, which can be thought as the value associated to a noun on a continuum, whose ends are lexical classes, and where polysemy is implicitly represented as an intermediate value between two classes. Such an index allows to represent sense modulation of disemous nouns (i.e. polysemes with two senses) and to account for the predominance of one the senses, if any, and therefore to accurately detect polysemes.

The main contribution of the work is a novel method for the identification and representation of polysemous nouns, which accounts for the semantic of such nouns and explicitly represents it.

2 Motivation and Related Works

In the field of NLP the information regarding lexical semantic classes has been proved to be crucial for several applications, such as information extraction, machine translation and question answering, and an increasing amount of research has been carried out in order to create models for the automatic classification of nouns (Romeo et al., 2014a; Bel et al., 2013; Schwartz et al., 2014).

Despite the effects of regular polysemy on lexical classification, few work attempted to computationally model the phenomenon. The approaches in the literature for the representation of polysemous words are basically three. Polysemes can be simultaneously represented as members of several classes (e.g. the polysemous word *xy* belongs to both the classes *X* and *Y*); as members of new, independent class which includes only words with the same hybrid distributional behaviour (*xy* belongs to the new class *XY*); on a continuum, thus assigning each word a polysemy index between 0 and 1.

Boleda et al. (2012a) present an in-depth study on adjective categorization with a special focus on polysemy, in which they conclude that multiple attribution is the best way to model polysemy. Boleda et al. (2012b) present a model to determine whether a noun matches a given sense alternation, while Romeo et al. (2013) introduce a supervised model for polysemy detection and conclude that since polysemes demonstrate lexico-syntactic traits of multiple classes, they can be considered as members of such classes. Romeo et al. (2014b) improve the previous approach reaching an accuracy of 60.71% in a polysemous noun classification task.

As pointed out by Boleda et al. (2012a) both the first two approaches are not completely satisfactory: the multi-labelling approach fails to represent the differences between polysemous and monosemous nouns in a class, while the second one does not account for the significant similarities between a polyseme and monosemous nouns belonging to the same classes of its senses. Furthermore, none of the two approaches can provide an adequate representation of the asymmetry of senses in a polysemous word.

The third approach has been explored by Utt and Padó (2011): they introduce a *polysemy index* based on the systematicity of sense variation of polysemous nouns to distinguish between polysemy and homonymy. Building on the same methodology, Frontini et al. (2014) define a threshold based on known basic type alternations, and use such a threshold to calculate the polysemy of new words. Finally, Martínez (2013) proposes an index that ranges from literal to metonymic, and that is used to account for the underspecified sense of polysemes.

Despite the fact that we also propose an index for polysemy representation, there are significant differences with the indexes proposed by the cited works. Martínez (2013) focuses on a particular aspect of polysemy, i.e. underspecification; Utt and Padó (2011) only take into account the distinction between homonimous and polysemous nouns; fi-

nally, Frontini et al. (2014) investigate the systematic aspect of the phenomenon, focusing on the detection of new basic sense alternations not considered in literature. Differently, in our work we aim to provide a single representation for each polysemous noun that explicitly accounts for its degree of membership to the basic lexical classes its senses belong to, thus highlighting the differences in distributional behaviour of different polysemes and addressing the problem of sense asymmetry.

3 Polysemy Detection

We present in this section a method for the identification of polysemous nouns. Firstly, a clustering algorithm was employed in order to verify that the semantic information preserved in WE was enough to separate nouns belonging to different lexical classes. Given a gold standard composed of *L* lexical classes and a set of nouns *N*=$\{n_1,n_2,...n_i\}$ distributed across these classes, WE representing the nouns in *N* were clustered in a number *L* of clusters. As a result of this first task, the clusters were expected to largely correspond to the classes in the gold standard.

Once the reliability of the semantic information in WE was proved, a second clustering task was performed to assess the following hypothesis: since polysemous nouns show distributional patterns that are different from non-polysemous nouns, WE representing polysemes are different from the others, and this difference can be captured by means of a clustering algorithm. This hypothesis was verified by clustering the WE of nouns in *N* in *L*+1 clusters: as a result, *L* clusters of monosemous nouns and one cluster composed of polysemes instantiating a specific alternation were expected. Note that the fact that just one class of polysemous nouns was expected only depended on the fact that all the cases of polysemy in the dataset instantiated a single sense alternation (animal/food).

3.1 Word Embeddings

WE of size 200 were trained using the word2vec toolkit[1] with the CBOW architecture, which has been proved computationally efficient for large datasets (Mikolov et al., 2013b; Baroni et al., 2014). Given the results of some preliminary studies, three relevant choices were made in the training phase. Firstly, WE were trained on a parsed version of the British National Corpus (BNC). The choice is in line with previous research (Levy and Goldberg, 2014) that proved how the embeddings created on input annotated with dependency relations better represent similarity (i.e. the paradigmatic relation existing between words, e.g. *coffee* and *tea*) compared with embedding created on linear contexts, which tend to encode more contextual information, or relatedness (e.g. *coffee* and *cup*).

Second, the size of the window was 1 word either side of the target word: once again, the reason is that smaller context windows have been proved to improve the ability of the model to represent similarity (Kiela and Clark, 2014).

Finally, consistently with the theoretic approach adopted, a single WE was created for each noun; thus, only one vector representation was available for the two senses of a disemous noun.

3.2 Clustering

WE have proved to be a representation that preserves semantic information in a vectorial space. Therefore, since nouns belonging to the same lexical class are close in the semantic space, a clustering algorithm should be capable of discovering the portion of the space where all the members of a class are located and include them in a cluster, thus separating them from nouns of other classes. We used the *k*-means algorithm, a flat, partitional algorithm that minimizes the distance from objects and their centroid and performs hard clustering.

3.3 Evaluation

For evaluation, we used the dataset proposed by Schwartz et al. (2014), which was built on the CSLB norms dataset created by Devereux et al. (2014), a very rich dataset made up of 638 concepts manually labelled by thirty annotators. Schwarz et al. (2014) applied a filtering mechanism to the original CSLB and obtained a final dataset made up of 346 nouns belonging to four semantic categories: animacy (*ANI*, 146 nouns), edibility (i.e. food items, *EDI*, 115 nouns), tools (*TOO*, 35 nouns) and things that can be worn (*WOR*, 50 nouns). The dataset included 33 disemous nouns which were represented

[1] https://code.google.com/p/word2vec/

cluster	k=4			k=4+1		
	precision	recall	f-score	precision	recall	f-score
0/WOR	0.98	0.92	0.95	1	0.92	0.96
1/EDI	0.83	0.78	0.81	0.87	0.94	0.90
2/ANI	0.85	0.85	0.85	0.89	0.95	0.92
3/TOO	0.77	0.97	0.86	0.81	0.97	0.88
4/ANI_EDI	/	/	/	0.83	0.44	0.65

Table 1: Results of the clustering tasks with k=4 and k=4+1.

as two lexical items (e.g. for the noun *chicken*, *chicken_ANI* and *chicken_EDI* [2]). The results of the first clustering task (k=4) were evaluated against this dataset.

For the second clustering (k=4+1), the dataset was slightly modified, and disemous nouns were represented with a new label indicating their polysemy (*chicken_ANI_EDI*).

In table 1 the results of the two clustering tasks are shown. For each cluster and its corresponding class in the dataset (first column) *precision*, *recall* and f-score for k=4 and k=4+1 were computed. The results confirmed that (i) the semantic information kept in WE was enough to include nouns of the same lexical class in the same cluster: this is proved by the high f-scores for k=4; (ii) WE preserved enough information to distinguish polysemous nouns instantiating a specific sense alternation from the other nouns in the dataset. As expected, when clustering with k=4+1, four out of five clusters were mainly composed of monosemous nouns, thus corresponding to the four classes *ANI*, *EDI*, *WOR* and *TOO* of the dataset, whereas the fifth included polysemes labelled as *ANI_EDI*. The f-score for this cluster was significantly lower than the others, but note that Martínez et al. (2013) report about the difficulties of the task also for human annotators.

The most relevant data are the precision and recall of the cluster *4/ANI_EDI*. We will discuss about the recall in the following section. The high result in precision confirmed that the great majority of the nouns in the cluster were polysemous, and hence that it was possible to find a portion of the semantic space populated only by polysemes.

Finally, the data regarding the f-scores with k=4+1 show that considering polysemous nouns as members of a new class led to an improvement of the results for all the other clusters. This confirms that by accounting for regular polysemy it is possible to improve the performance of a system in a task of lexical classification.

3.4 Discussion

From the error analysis, two general errors causes and a specific one for k=4 and k=4+1 were identified. Firstly, the *association effect*: even if the setting chosen for the creation of the WE was intended to maximize similarity and to minimize relatedness, a noun like *peeler_TOO*, which co-occurs almost exclusively with nouns referring to fruits and vegetables (akin nouns belonging to *EDI*), was included in *1/EDI* instead of *3/TOO*. The same effect has also been found in other works (Hill et al., 2014).

Secondly, the *low frequency effect*: the majority of the misclassified monosemous nouns had less than 50 occurrences in the BNC corpus (e.g. *chipmunk_ANI*, 20 occurrences). Few occurrences of a word seemed to produce less informative WE and this, in turn, less accurate cluster assignment. Note that WE for low frequency words are usually discarded (Mikolov et al., 2013c).

For the first task (k=4) the *polysemy effect*: 36 out of 51 errors were due to regular polysemy, an expected result, given that only one embedding for each noun had been used. Thus, for example, since *chicken_ANI* and *chicken_EDI* were represented by the same vector, they were both assigned to *1/EDI*.

Finally, the low recall of the cluster for polysemous nouns in k=4+1 was due to the already mentioned sense asymmetry: when one sense is more

[2] The notation '*_CAT*' is used to identify the class in the dataset which a lexical item belongs to, while '*_n/CAT*' identifies the cluster - and the corresponding class - to which a noun is assigned by our method.

taxonomic labelling	production frequency	taxonomic labelling	production frequency
crustacean	6	edible_eaten	20
shellfish	7	seafood	6
tot animal	13	tot food	26
% ANI	33	% EDI	67

Table 2: Taxonomic labelling and production frequency for the noun *prawn*

cluster	% EDI-related human judgements	% ANI-related human judgements
1/EDI	64	36
2/ANI	33	67
4/ANI_EDI	48	52

Table 3: Averaged taxonomic labelling for nouns in *1/EDI*, *2/ANI* and *4/ANI_EDI*

frequent or predominant that the other, the WE is expected to be more similar to those of non ambiguous nouns in a basic class, as it mostly behaves, in distributional terms, like them.

Sense predominance was also observed in the human annotated CSLB dataset. For every concept in the CSLB, information about the taxonomic group the concept belongs to is also provided as well as the relative production frequency, i.e. the number of times the taxonomic group has been associated to the target concept. As an example, for the noun *prawn* the information reported in table 2 is provided. Our assumption was that the difference in taxonomic labelling adequately reflects the difference in sense predominance: hence *prawn*, on average, is associated to the class *EDI* on 67% of times, and to *ANI* in the remaining 33%.

In table 3, we averaged relevant values of production frequency for polysemous nouns in *4/ANI_EDI*, *2/ANI* and *1/EDI*. On average, polysemous nouns that in task *k*=4+1 were included in cluster *1/EDI* were judged to belong to the taxonomic category of food in 64% of cases, and only 36% to the class of animals; polysemous nouns assigned to cluster *2/ANI*, were labelled by humans as animals in 67% of cases (33% food). Finally, polysemous nouns clustered in *4/ANI_EDI* had almost equal average of human labelling for the two senses: 48% food, 52% animal.

Human judgements support our explanation of the low recall of cluster *4/ANI_EDI*: on average, the same polysemous nouns that were incorrectly included in *1/EDI* and *2/ANI* because of their strong sense asymmetry, are mostly considered as food or animal respectively by humans.

On the basis of these data, it is possible to conclude that, as expected, cluster *4/ANI_EDI* included only polysemous nouns, but only those whose senses are *balanced*, i.e. not strongly asymmetric. On the contrary, *unbalanced* polysemous nouns tended to be included in the classes the predominant sense belongs to.

4 Modeling Sense Asymmetry

The clustering introduced above allowed to identify most polysemous nouns for a specific class alternation but failed to handle polysemes with a strong sense asymmetry. In order to overcome this problem, in this section we propose a *sense index* which, given a specific class alternation, makes it possible to identify and represent disemous nouns and their particular sense asymmetries. Such an index allows to detect and accurately represent polysemes, since it locates nouns on a continuum whose ends are the two involved lexical classes: monosemous nouns lay close to one of the ends, while polysemous nouns lay in the middle of the continuum, in a different position depending on the degree of asymmetry of their senses.

The sense index was calculated considering the *k*=4 scenario, that is without taking into account the polysemous class *ANI_EDI*. For each cluster, the centroids *C* were computed. A centroid is an average vector and hence, in our method, it represents the centre of a class; therefore, the closer a noun is to the centroid of a class, higher its degree of membership to such a class is.

Our goal was to identify which nouns were polysemous for a specific class alternation; for this reason the method considered one pair of clusters *X*, *Y* at a time. Given *X*=$\{x_1, x_2,...x_n\}$ and *Y*=$\{y_1, y_2,...y_n\}$, and the centroids of the two addressed clusters C_x and C_y, the cosine similarity between each centroid and every noun in the two classes was computed to assess class membership. In what fol-

lows, we will refer to this initial value as *absolute cosine class membership* (*θAbs*).

The highest (*θMaxAbs*) and the lowest (*θMinAbs*) values of *θAbs* were then used in order to define a *relative cosine class membership* (*θRel*). The reason for switching from absolute to relative class membership is the following. WE represent the whole vocabulary of the corpus in a single vector semantic space: since *θAbs* is computed in this space, it accounts for the distance between the nouns of the dataset and all the other words of the vocabulary. This fact had a major drawback: while accounting for the cosine distance with very distant words in the semantic space, it minimized the distances between class-related nouns. By adopting *θRel* we overcame this problem, since this measure is aimed to only account for the semantic space in which nouns of two classes and their centroids are included, thus making evident the differences of targeted nouns.

Relative cosine class membership between a centroid and a noun *n* in the dataset can be computed as in (1):

$$\theta Rel = \frac{\theta Abs - \theta MinAbs}{RanMaxMin} \quad (1)$$

where *θAbs* is the absolute cosine similarity between the target name and the centroid, *θMinAbs* and *θMaxAbs* are the values of cosine similarity between the centroid and the farthest and closest point respectively, and *RanMaxMin* is the range between these two values.

Let's consider the noun *fox* and the centroids C_{ani} and C_{edi}. The *θAbs* between *fox* and C_{ani} was 0.73, between *fox* and C_{edi} 0.48, while the values of *θRel* were 0.80 with C_{ani} and 0.30 with C_{edi}. Clearly, relative cosine similarity made more evident the proximity in the semantic space of *fox* to C_{ani}, and stressed the distance between *fox* and C_{edi}.

Finally, the values of *θRel* between a noun and the two centroids were used to obtain the sense index, whereby the degree of membership of a noun to two classes is defined. Given the two values of relative cosine class membership between a noun and the two centroids θRel_1 and θRel_2 sense index for the first sense was computed as in (2):

$$SenseIndex1 = \frac{\theta Rel_1{}^2}{\theta Rel_1{}^2 + \theta Rel_2{}^2} \quad (2)$$

Since the indexes for the first and the second sense are complementary, the second index was obtained by computing 1-SenseIndex1. For *fox* the final sense index was 0.88 for *ANI* and a corresponding 0.12 *EDI*, a result that indicates a clear membership of the target noun to lexical class of the animals, and therefore that *fox* can be considered a monosemous noun. The differences between monosemous and polysemous are evident for nouns in table 4.

target noun	dataset label	ANI-index	EDI-index
butter	EDI	0.08	0.92
bacon	EDI	0.07	0.93
cheese	EDI	0.08	0.92
eagle	ANI	0.87	0.13
panther	ANI	0.95	0.05
pelican	ANI	0.98	0.02
calf	ANI_EDI	0.51	0.49
shrimp	ANI_EDI	0.40	0.60
chicken	ANI_EDI	0.31	0.69

Table 4: Examples of sense indexes

The indexes in the table allow to distinguish three groups of nouns, a finding which is in line both with those of the clustering method and with the initial hypothesis that polysemous nouns gather together in a specific area of the vectors semantic space. In the first part of the table monosemous nouns labelled as *EDI* in the dataset are listed: their sense index clearly reflects this class attribution, as all the indexes for *EDI* are above 0.90. The same holds true for the nouns in the second part of the table, i.e. those nouns with high *ANI* index. The third part of the table is the most interesting one. Firstly, the polysemy of the nouns in this section can be easily detected by looking at their balanced indexes. Second, the method provided specific information that stresses the differences regarding the distributional behaviour of polysemes: for example, while with the clustering method the only available information was that *calf* and *shrimp* were in the same cluster (*ANI_EDI*), with the present method we know that while the two senses of *calf* are almost perfectly

balanced (0.51 *ANI*/0.49 *EDI*), *shrimp* is more similar to the monosemous nouns in *EDI* (0.60) than to those in *ANI* (0.40). Finally, the present method overcame the limitation of the previous one related to sense asymmetry. For example, *chicken* was in the cluster *1/EDI* in *k*=4+1 given its strongly predominant food sense, while its sense index is 0.69 *EDI*/0.31 *ANI*: by comparing this index with that of the monosemous nouns of the first part of the table 4, it becomes clear that *chicken* is not so strictly tied to the food sense, i.e. it is not a monosemous noun, but also that, differently from other polysemous nouns in the third part of the table, its *EDI*-related sense is strongly predominant.

4.1 Evaluation

In order to asses the quality with which our method distinguished between polysemous and monosemous nouns we turned the evaluation of the sense index in a classification task. For this purpose, given a pair of lexical classes, we identified two thresholds 'α' and 'γ' on the continuum between such classes, thus obtaining three separate ranges. For such ranges, we expected the following distribution: monosemous nouns labelled as the first sense in the range [0-α], nouns labelled as the second sense in the range [γ-1] and polysemous nouns instantiating the alternation of the two senses in [α-γ]. The populations of the three ranges were than evaluated against the classes of the dataset annotated with polysemous nouns, from which we removed WE that proved to preserve insufficient information due to low frequency of the noun in the corpus (see 3.4.). We selected two pairs of lexical classes for the evaluation, namely *ANI-EDI*, a typical regular polysemy alternation, and *WOR-EDI*, an impossible alternation according to the literature. While for the former a certain number of polysemes in the middle of the continuum was expected, for the latter the expectation was to find two groups of nouns laying at the ends of the continuum, and nothing in the middle.

Along with Utt and Padó (2011), in order to chose the best threshold for the two pairs of classes we experimented with different values for α and γ and evaluated the resulting populations of the ranges with the Mann-Whitney U-Test, a non parametric test that is used to test whether two populations are significantly different or not. The output of the test is the *U*-value, which can range from 0 to a number computed considering the values of the two populations in exam. In our case, values of *U* close to 0 would mean that monosemous and polysemous nouns lay in different ranges, while high values of *U* indicate that they are approximately evenly distributed on the continuum.

For *ANI-EDI* we identified the best threshold for α=0.23 (*U*=46 on a maximum value of 5671) and γ=0.59 (*U*=227, max value 9730), with p$\leq$0.05 for both the results. As expected, *EDI*-nouns were in the range [0-α], ANI-nouns in the range [γ-1] and polysemes in [α-γ]. The low values of *U* indicate that, even if there was a partial overlapping of the three populations on the continuum, there were significant differences among them.

For *WOR-EDI* the identification of best values for α and γ was fairly straightforward. Since all the nouns labelled as *WOR* were in the range [0-0.26] and the nouns labelled as *EDI* in the range [0.63-1], the values of the two thresholds were at α=0.26 (*U*=0, p$\leq$0.05) and γ=0.63 (*U*=0, p$\leq$0.05). In table 5 are shown the results obtained using the polysemy index for the classification of polysemous nouns for the two class pairs *ANI-EDI* and *WOR-EDI*.

	ANI-EDI		
range	precision	recall	f-score
[0-α]/EDI	0.97	0.96	0.96
[α-γ]/ANI_EDI	0.73	0.88	0.81
[γ-1]/ANI	0.97	0.93	0.95
	WOR-EDI		
range	precision	recall	f-score
[0-α]/WOR	1	1	1
[α-γ]/WOR_EDI	/	/	/
[γ-1]/EDI	1	1	1

Table 5: Results of the evaluation of the Sense Index.

4.2 Discussion

For both the class alternations taken into account, the results confirmed our expectation to find monosemous nouns on the ends of the continuum, and disemous nouns (if present) in the middle. This was especially true for *WOR-EDI*, in which nouns were clearly polarized at the ends of the continuum and no polysemous nouns instantiating this alterna-

tion were found among them.

Following the comparison of the sense index results for *ANI-EDI* with those of the *k*=4+1 clustering, some conclusions can be drawn. Firstly, the general improvement in the f-score of *ANI*, *EDI* and *ANI_EDI* proved that the method presented in this section outperformed the clustering algorithm in the task of polysemy detection.

Secondly, the significant improvement in recall for *ANI_EDI* showed that by exploiting the information provided by the sense index it was possible to detect those polysemous nouns with a strongly predominant sense that were misclassified with the previous method, even if there were few exceptions. *Tuna* was found among nouns laying in the *EDI* range, but close to the threshold α (*tuna*'s index is 19, α=23). As for polysemes found in the *ANI* range some explanations can be advanced: *sheep* and *cow* are not usually employed to denote meat as they have a corresponding term specific for this use; *snail* is very seldom meant as food; for *rabbit* we conjecture that there are a number of figurative uses that made it more distant for the group of edible things.

The results show that the sense index has considerably increased *ANI_EDI* recall but, to some extent, at the expenses of the precision, what was already expected after the *U*-test which showed the existence of some overlap between ranges. Nevertheless, the general increase in f-score terms for the three involved classes shows the benefits of the index approach.

The correspondence of the values of the thresholds α and γ placed on the two continua was almost perfect: 0.23 and 0.59 on *ANI-EDI*, 0.26 and 0.63 on *WOR-EDI*. This means that the separation of the ranges on the continuum was consistent between different senses alternations, thus proving the robustness of the method.

Finally, the considerable distance between the first polysemous noun (*tuna*), laying at point 0.19, and the last one (*octopus*) at point 0.70 on the *ANI-EDI* continuum, reflects the fact that the distributional characteristics of polysemes are more disperse than those of monosemous nouns.

5 Conclusions

We have presented an ongoing work that proposes a method for the detection and representation of polysemous nouns based on the semantic information preserved in WE. We initially showed that polysemous nouns instantiating a particular sense alternation group together in a specific area of the vector semantic space. Subsequently, we proposed a method by which, given a pair of lexical classes, a sense index defining the degree of membership to such classes is assigned to each noun. In this method, polysemy is implicitly represented as a balanced value of the degree of membership. We finally showed that it is possible to identify two thresholds α and γ such that nouns having sense index included in the ranges [0-α] and [γ-1] are monosemous nouns belonging to the first and the second class of the pair respectively, while polysemes are included in the range between α and γ.

Results, although limited in scale, show that the method allows a clear separation between monosemous belonging to a lexical class and polysemous nouns instantiating specific sense alternation, also accounting for those polysemes with a strong asymmetry in sense predominace.

In future work, we plan to double check that the results are independent of the datasets, provided enough data is considered. Furthermore, more experiments with other class alternations are planned as well as particular application of our sense index in actual WSD applications.

6 Acknowledgements

This work was funded with the support of the IULA-UPF PhD fellowship program.

References

Ju D Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Núria Bel, Lauren Romeo, and Muntsa Padró. 2013. Au-

tomatic lexical semantic classification of nouns. *arXiv preprint arXiv:1303.1930*.

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

Gemma Boleda, Sebastian Padó, and Jason Utt. 2012b. Regular polysemy: A distributional model. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 151–160. Association for Computational Linguistics.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of semantics*, 12(1):15–67.

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

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Francesca Frontini, Valeria Quochi, Sebastian Padó, Jason Utt, and Monica Monachini. 2014. Polysemy index for nouns: an experiment on italian using the parole simple clips lexical database.

Felix Hill, KyungHyun Cho, Sebastien Jean, Coline Devin, and Yoshua Bengio. 2014. Not all neural embeddings are born equal. *arXiv preprint arXiv:1410.0718*.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, pages 21–30.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Héctor Martínez Alonso, Bolette Sandford Pedersen, and Núria Bel. 2013. Annotation of regular polysemy and underspecification. In *ACL (2)*, pages 725–730.

Héctor Martínez Alonso. 2013. Annotation of regular polysemy: an empirical assessment of the underspecified sense.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013d. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Geoff Nunberg. 1992. Systematic polysemy in lexicology and lexicography.

James Pustejovsky. 1991. The generative lexicon. *Computational linguistics*, 17(4):409–441.

J Pustejovsky. 1995. Tile generative lexicon. *ms. Brandeis University*.

Lauren Romeo, Sara Mendes, and Núria Bel. 2013. Towards the automatic classification of complex-type nominals. In *6th International Conference on Generative Approaches to the Lexicon*, page 21. Citeseer.

Lauren Romeo, Gianluca E Lebani, Núria Bel, and Alessandro Lenci. 2014a. Choosing which to use? a study of distributional models for nominal lexical semantic classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4366–4373.

Lauren Romeo, Sara Mendes, and Núria Bel. 2014b. A cascade approach for complex-type classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4451–4458.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2014. Minimally supervised classification to semantic categories using automatically acquired symmetric patterns. In *Proceedings of COLING 2014 the 25th International Conference on Computational Linguistics*, pages 1612–1623.

Jason Utt and Sebastian Padó. 2011. Ontology-based distinction between polysemy and homonymy. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 265–274. Association for Computational Linguistics.