First-order and second-order context representations: geometrical considerations and performance in word-sense disambiguation and discrimination

Alfredo Maldonado-Guerra and Martin Emms
maldonaa@scss.tcd.ie, mtemms@scss.tcd.ie

School of Computer Science and Statistics
Trinity College Dublin
Ireland

Abstract

First-order and second-order context vectors ($\mathbf{C}^1$ and $\mathbf{C}^2$) are two rival context representations used in word-sense disambiguation and other endeavours related to distributional semantics. $\mathbf{C}^1$ vectors record directly observable features of a context, whilst $\mathbf{C}^2$ vectors aggregate vectors themselves associated to the directly observable features of the context. Whilst $\mathbf{C}^2$ vectors may appeal on a number of grounds, such as being less sparse and leveraging additional information from a larger corpus, not much work has been devoted to contrasting $\mathbf{C}^2$ with $\mathbf{C}^1$ vectors. While the concerns of the paper are primarily empirical we also advocate a particular formulation of $\mathbf{C}^2$ vectors, whereby $\mathbf{C}^2$ vectors (of dimensionality f_2) are derived from $\mathbf{C}^1$ vectors (of dimensionality f_1) by post-multiplication by some $f_1 \times f_2$ matrix. This makes plainer the relation of the $\mathbf{C}^2$ construction to standard methods for dimensionality reduction. We then consider two geometric properties of $\mathbf{C}^1$- and $\mathbf{C}^2$-based sense vectors for sense-tagged data. We show that, perhaps surprisingly, the $\mathbf{C}^2$-based representation of a sense is not to any great extent parallel (similar) to the $\mathbf{C}^1$-based representation of that sense. We also show that the angular spread amongst the $\mathbf{C}^1$-based sense vectors is considerably greater than the spread amongst the $\mathbf{C}^1$ versions. Following on from this, we then compare both sense vectors in supervised word sense disambiguation and unsupervised word sense discrimination settings, finding the $\mathbf{C}^1$-based vectors superior to the $\mathbf{C}^2$-based vectors in the supervised setting, but quite similar in performance in the unsupervised setting.

Keywords: first-order context vectors, second-order context vectors, word-sense disambiguation, word-sense discrimination, distributional semantics

1. Introduction

In computational lexical semantics, the so-called *Vector Space Model* (Salton, 1971; Turney and Pantel, 2010) assumes that word meanings, and relations amongst them, can be modelled with vectors, and geometrical notions based on them. The dimensions of these vectors are typically

This research is supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin. All calculations were performed on the Lonsdale cluster maintained by the Trinity Centre for High Performance Computing. This cluster was funded through grants from Science Foundation Ireland.

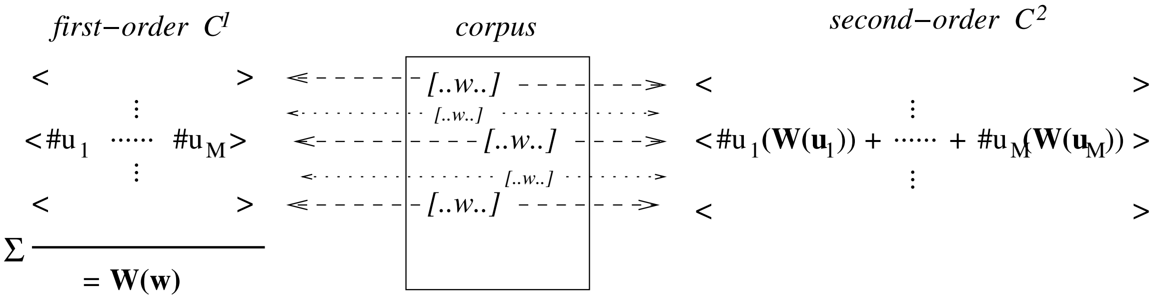

Figure 1: To the left $\mathbf{C^1}$ *vectors for different occurrences of* w *–* $\#u_i$ *is count of unigram* u_i *in a window around the occurrence of* w*. Summing all gives* $\mathbf{W}(w)$*. To the right, the* $\mathbf{C^2}$ *vectors, summing the* $\mathbf{W}(u_i)$ *of the unigrams* u_i *in the window.*

identified with directly observable aspects of word occurrences, and in the simplest and most widely adopted approach the dimensions are identified with *unigrams* that co-occur within a certain window of a target word. For a *token* of word w, or equivalently a position p in a document, one can define what is sometimes termed the **first-order context vector** of the instance p, $\mathbf{C^1}(p)$, such that $\mathbf{C^1}(p)[u]$ – the value of the vector $\mathbf{C^1}(p)$ for the unigram u – simply counts how many occurrences of u there are within a specified distance of the particular occurrence of w at p. From these *token* representations, for a word *type*, one can define what is often termed the **word vector**, $\mathbf{W}(w)$, by some kind of aggregation over the token representations, and in the simplest case, this is a sum. The result is that for a unigram u, $\mathbf{W}(w)[u]$ counts how many occurrences of u there are within a specified distance of *any* occurrence of w in the corpus — see left-hand side of Figure 1. Similarity of such word vectors[1], as quantified for example by cosine, has been used with some success as a means of assessing similarity of word meaning (Lin, 1998; Rapp, 2003). The same approach can be adopted to seek to model the distinct senses of an ambiguous word. Thus suppose w has K senses $s_1 \dots s_K$. If $\mathcal{P}_i$ is the set of w's occurrences manifesting sense s_i, then by aggregating the vectors $\mathbf{C^1}(p)$ over just these occurrences, you obtain what can be termed the **first-order sense vector**, $\mathbf{S^1}(\mathcal{P}_i)$; the word-vector is just the special case in which *all* occurrences of the word are counted. Such vectors have been used for word-sense disambiguation (Oh and Choi, 2002; Sugiyama and Okumura, 2009; Martinez and Baldwin, 2011).

In *unsupervised* word-sense induction or discrimination tasks the aim is to induce a sense-reflecting partition over the tokens of an ambiguous word. It was working in this unsupervised setting that Schütze (1998) introduced so-called **second-order** variants of context vectors and sense vectors. A **second-order context vector**, $\mathbf{C^2}(p)$, represents an instance p not directly in terms of its neighbouring unigrams, but instead takes the *word-vectors* of those neighbouring unigrams and sums them — see Figure 1, right. Because the word-vector of a neighbouring unigram u contains counts for all the unigrams v with which u co-occurs, a value $\mathbf{C^2}(p)[v]$ in the second-order vector is said to refer to second-order co-occurrence information.

Just as a $\mathbf{S^1}$ vector can be defined from the $\mathbf{C^1}$ vectors for a particular set of occurrences, Schütze (1998) defines a **second-order sense vector**, $\mathbf{S^2}$ by aggregating the $\mathbf{C^2}$ vectors for a particular set of occurrences. A number of other authors have since worked with essentially these second-order representations (Purandare and Pedersen, 2004; de Marneffe and Dupont, 2004; Sagi et al.,

[1] or transforms of them, converting counts to measures of association

2008; Wang and Hirst, 2010).

Thus there are two contending vector representations of a word's sense, $\mathbf{S^1}$ and $\mathbf{S^2}$, differing simply according to whether they aggregate $\mathbf{C^1}$ or $\mathbf{C^2}$ vectors. There are diverse intuitions/motivations for using $\mathbf{S^2}$ and $\mathbf{C^2}$ vectors, rather than their first-order counterparts. One intuition is that $\mathbf{C^1}$ vectors will tend to be *sparse*, having zeros on all but a small number of dimensions and that $\mathbf{C^2}$ vectors will tend to be less so. A related intuition is that a $\mathbf{C^2}$ vector brings additional information from words co-occurring elsewhere with the context's own words, potentially in a much larger corpus than the corpus on which word-sense discrimination is being done. These are intuitions only and there has not been a great deal of work systematically comparing the representations. Purandare and Pedersen (2004) compare (a certain variant of) $\mathbf{C^1}$ vectors with (a certain variant of) $\mathbf{C^2}$ vectors, leading to a conclusion that $\mathbf{C^1}$ out-performed $\mathbf{C^2}$ when there are many examples ($\approx 4,000$), but $\mathbf{C^2}$ out-performed $\mathbf{C^1}$ when there were few examples (a few hundred). The aim of this paper is to carry out a more systematic comparison of these representations than, we believe, has hitherto been done.

Section 2 formally defines the necessary notions. In doing this, we provide an alternative but equivalent to the usual derivation of second-order context vectors, whereby a $\mathbf{C^2}$ vector (of dimensionality f_2) is derived from a $\mathbf{C^1}$ vector (of dimensionality f_1) by post-multiplication by a $f_1 \times f_2$ matrix. This makes plainer the relation of the $\mathbf{C^2}$ construction to SVD methods for dimensionality reduction. Before evaluating first-order and second-order representations on particular tasks, section 3.2 considers the *geometry* of sets of $\mathbf{S^1}$ vectors and $\mathbf{S^2}$ vectors. One question looked at is the extent to which the $\mathbf{S^1}$ and $\mathbf{S^2}$ vectors for a given sense are approximately *parallel* to each other. Another question looked at is whether the $\mathbf{S^1}$ vectors for the different senses of a given word show a comparable angular spread amongst themselves to the corresponding $\mathbf{S^2}$ vectors. Section 3.3 then evaluates first-order and second-order representations in *supervised* word-sense disambiguation experiments, whilst section 3.4 compares them in *unsupervised* word-sense discrimination experiments. Section 4 then describes previous related work, and draws conclusions.

2. Definitions

Let doc be a corpus, and let $win^l(p)$ be a function returning a set of positions p' around p – the 'window' around p. l is the window width[2] and typically $p' \in win^l(p)$ iff $p-l/2 \leq p' \leq p+l/2$. Then in general a *feature* can be identified with a function that maps windows to $\mathbb{R}$. For a particular position p in doc (or equivalently a particular *token* of a word), the **first-order context vector**, $\mathbf{C^1}(p)$, is a vector giving the values of the features in the window around p. In most cases, features are equated with *unigrams*, one for each member of Σ_f, some chosen subset of the unigrams of the corpus doc[3], and the value of a unigram feature u on the window $win^l(p)$ is simply the count of u in the window:

Definition 1 (First-order context vector (unigrams))**.** *For window width l, and a choice of dimensionality f corresponding to a restriction to some unigram vocabulary Σ_f, the first-order context vector for position p, $\mathbf{C^1}(p)$ is the vector of dimensionality f such that for any $u \in \Sigma_f$, $\mathbf{C^1}(p)[u] =$ frequency of u in $win^l(p)$.*

[2] In this work, we set $l = 20$ for all experiments.

[3] This subset of unigrams Σ_f is a selection of f words from the corpus that fulfil some criteria. Following Schütze (1998), experiments in this work select the top f most frequent words in the corpus, excluding function (stop) words, but many other selection criteria are possible.

A **word vector** is a vector to represent a particular word *type* w in a corpus, and is based on the set of $\mathbf{C^1}$'s for its *tokens*. The simplest possibility is just to *sum* these.

Definition 2 (Word vector)**.** *Assuming* $\mathbf{C^1}(p)$ *is defined for all positions* p*,*

$$\mathbf{W}(w) = \sum_{p:\mathrm{doc}[p]=w} (\mathbf{C^1}(p)) \tag{1}$$

With unigram features, on this definition $\mathbf{W}(u)[v]$ is simply the count of how often u and v co-occur in the corpus within $l/2$ words of each other, where l is the chosen window-width for the $\mathbf{C^1}$ vectors.

As already noted, Schütze (1998) suggested an alternative **second-order context vector**, $\mathbf{C^2}(p)$, for a particular position p , or equivalently, particular *token* of a word. In its simplest incarnation, it is simply *the sum of the word vectors of words in the window*

$$\mathbf{C^2}(p) = \sum_{p' \in win^l(p), p' \neq p} (\mathbf{W}(doc[p'])) \tag{2}$$

This $\mathbf{C^2}$ notion has to be elaborated a little further, but first let us define first and second order *sense vectors*. If w is an ambiguous word and $\mathcal{P}$ is a set of positions exhibiting a particular sense, then the centroid of the context-vectors for the occurrences in $\mathcal{P}$ is a candidate representation of the sense:

Definition 3 (First- and second-order sense vector, $\mathbf{S^1}$, $\mathbf{S^2}$)**.** *If* $\mathcal{P}$ *is a set of positions, the first-order sense and second-order sense vectors based on* $\mathcal{P}$ *are* $\mathbf{S^1}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{C^1}(p)$ *and* $\mathbf{S^2}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{C^2}(p)$

In (2), every $p' \in win^l(p)$ is summed over, but it is natural to consider restricting the sum to positions featuring words in a specified vocabulary, a vocabulary which need not coincide with the vocabulary used by the word vectors themselves: in Schütze (1998), it does not[4]. As a step towards addressing this, define first **WM** as the matrix of all word vectors:

Definition 4 (Word Matrix)**.** *Let* **WM** *be a* $f_1 \times f_2$ *matrix, with* f_1 *the size of some chosen vocabulary* Σ *and* f_2 *the size of the feature set used by the word vectors. The* i^{th} *row of* **WM** *is the word vector for the* i^{th} *word in* Σ*.*

A $\mathbf{C^2}$-definition parametrised by a word-matrix can then be given:

Definition 5 (second-order context vector (with parameter **WM**))**.** *Given a* $f_1 \times f_2$ *word matrix* **WM***, of word vectors for words in some chosen vocabulary* Σ *of size* f_1*, if* $\mathbf{C^1}(p)$ *is the* f_1*-dimensional first-order context vector representation of position* p *(using* Σ *for features), then* $\mathbf{C^2}(p)$ *the* f_2*-dimensional second-order context vector representation of position* p *is defined by*

$$\mathbf{C^2}(p) = \mathbf{C^1}(p) \times \mathbf{WM} \tag{3}$$

[4] He sums word vectors for the top 20,000 most frequent words (excluding function words) and uses the top 2,000 most frequent words (excluding function words) as features of these. So, in our notation we say that Schütze (1998) specifies dimensionalities of $f_1 = 20,000$ and $f_2 = 2,000$.

This formulation of the $\mathbf{C^2}$ construction is not the customary one, but we think a useful one. Although (3) and (2) look different, (3) is also defining $\mathbf{C^2}(p)$ to be a sum of multiples of relevant rows of **WM**, and in fact, if f_1 is defined so that there is a row of **WM** for *every* word in $win^l(p)$, it is easy to see that then (3) and (2) define the same vectors. Moreover, it is not hard to show that the relation between $\mathbf{C^1}$ and $\mathbf{C^2}$ vectors exemplified in (3) lifts to a relation between $\mathbf{S^1}$ and $\mathbf{S^2}$, given the definition of sense vectors as averages over $\mathbf{C^1}$ and $\mathbf{C^2}$ vectors:

$$\mathbf{S^2}(\mathcal{P}) = \mathbf{S^1}(\mathcal{P}) \times \mathbf{WM} \tag{4}$$

Because the word vectors on the rows of **WM** are typically shorter than the height of **WM** i.e. $f_2 < f_1$, equation (3) typically represents a dimensionality lowering transformation. It is interesting to note, that formulating the $\mathbf{C^2}$ construction as we have, (3) is strikingly similar to the defining equations for dimensionality reduction via Singular Value Decomposition (SVD). If $\boldsymbol{A}$ is a $M \times N$ matrix, it has a so-called 'reduced-rank SVD' $\hat{\boldsymbol{A}} = \boldsymbol{U}_k \times \boldsymbol{S}_k \times (\boldsymbol{V}_k)'$ (see Manning et al. (2008)), where amongst other things, $\boldsymbol{V}_k$ has the first k eigenvectors of $A \times A'$ for columns. Via the SVD, an N dimensional vector $\boldsymbol{t}$ is projected to a k-dimensional $\hat{\boldsymbol{t}}$ by

$$\hat{\boldsymbol{t}} = \boldsymbol{t} \times \boldsymbol{V}_k \tag{5}$$

Applied to any $\boldsymbol{A}$ whose rows have same dimensionality as that of $\mathbf{C^1}$ vectors, the SVD dimensionality reduction procedure would lead via (5) to a transformation of a $\mathbf{C^1}$ vector by post-multiplying by $\boldsymbol{V}_k$, a matrix of dimensions $f_1 \times k$. The $\mathbf{C^2}$ construction as defined in (3) also post-multiplies $\mathbf{C^1}$ by matrix **WM**, the $f_1 \times f_2$ word-matrix.

Although the definition of $\mathbf{C^2}$ can be elaborated further, these definitions should suffice for the experiments that follow.

3. Experiments

3.1. Corpora

The experiments to be reported make use of two datasets. One is the *hard-interest-line-serve* dataset (henceforth called "the HILS dataset") widely used in the word-sense disambiguation literature[5]. For each target word it contains sense-tagged short context samples from newspaper articles written between 1987 and 1991. See Figure 2 for the number of instances of each target word and their sense distribution. The second dataset is a larger corpus of untagged articles from the New York Times (NYT) from the 1998-2000 period[6]. The NYT dataset consists of 1.92×10^8 tokens distributed in 205990 articles.

The untagged NYT dataset is used in several different ways. We use the *pseudoword* technique (Yarowsky, 1993) whereby occurrences of two unrelated words (for example *banana* and *moon*) are replaced by their concatenation (*banana_moon*), creating a resolution task to revert each pseudoword correctly to the word replaced. Applied carefully, this generates a larger number of training and test instances than in a sense-tagged corpus. We use the pseudowords introduced by de Marneffe and Dupont (2004). We list them here with their total number of occurrences and their constituent distribution in the NYT: *animal_river* (total: 11808, *animal*: 33%, *river*: 67%), *banana_moon* (total: 3953, *banana*: 24%, *moon*: 76%), *data_school* (total: 49498,

[5] This dataset is in the public domain and can be freely downloaded from http://www.d.umn.edu/~tpederse/data.html

[6] The NYT articles are part of the AQUAINT Corpus:
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T31

TARGET	*hard*	TARGET	*interest*	TARGET	*line*	TARGET	*serve*
INSTANCES	4333	INSTANCES	2368	INSTANCES	4146	INSTANCES	4378

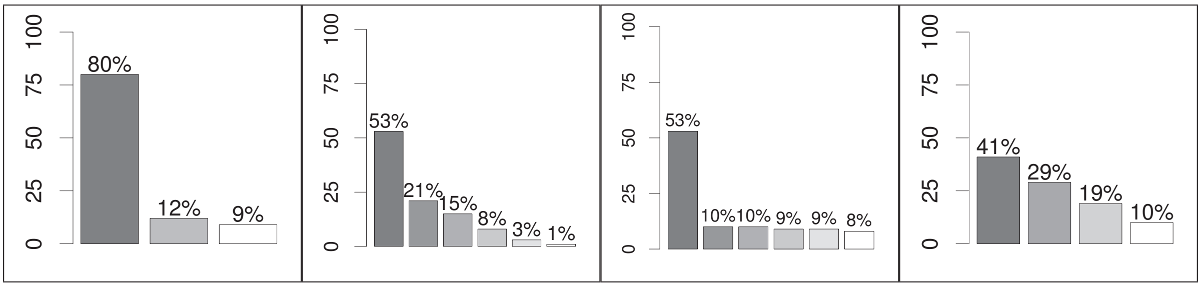

Figure 2: HILS dataset sense distributions

data: 21%, *school*: 79%), *railway_admission* (total: 3733, *railway*: 14%, *admission*: 86%) and *rely_illustration* (total: 4970, *rely*: 78%, *illustration*: 22%). In work with $\mathbf{C}^2$ vectors, the NYT dataset is also a possible source of the word vectors used in their construction. Finally, it is possible to perform unsupervised experiments on the untagged NYT and evaluate this via the HILS set, as will be seen in section 3.4.

3.2. Geometric properties of $\mathbf{C}^1$ *and* $\mathbf{C}^2$ *vectors*

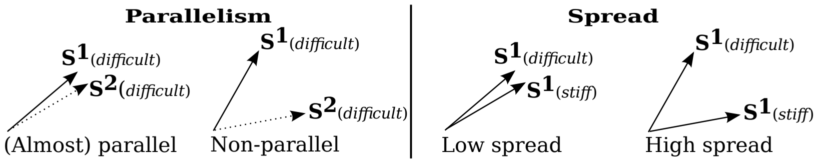

Figure 3: A depiction of parallelism and spread in sense vectors representing two senses for ***hard*** *– On the left-hand side, we measure how parallel different-order sense representations of the same sense are between each other. On the right-hand side, we measure how spread same-order sense vectors representing different senses are between each other.*

Since the general philosophy of the *Vector Space Model* is to model meanings with vectors, and to model semantic relations with geometrical notions based on these vectors, it is reasonable to consider ways to compare the geometry of sets of modelled meanings under the first-order and second-order approaches.

Parallelism If $\mathcal{P}_i$ is all positions instantiating a particular sense s_i of an ambiguous item, one can take the alternative sense vectors $\mathbf{S}^1(\mathcal{P}_i)$ and $\mathbf{S}^2(\mathcal{P}_i)$ and assess how *parallel* they are to each other, by computing their cosine. The left-hand side of Figure 3 shows two examples of how parallel an $\mathbf{S}^1$ vector and an $\mathbf{S}^2$ vector can be, both representing the sense *difficult* of the adjective *hard*. The more parallel (and therefore more similar) they are, the more they are approximations of each other.

To ensure the $\mathbf{C}^1$ and $\mathbf{C}^2$ vectors (and thereby derived $\mathbf{S}^1$ and $\mathbf{S}^2$ vectors) have the same dimensions, we use unigrams for their features, from a vocabulary Σ_{f_2}, and build a *square* $f_2 \times f_2$ word-matrix. We can then make f_2-dimensional $\mathbf{C}^1$ vectors, which convert to further f_2-dimensional $\mathbf{C}^2$ vectors (see definition 5). For the HILS data, the word-matrix used in the derivation of $\mathbf{C}^2$ vectors was computed once in a *local* fashion and once in a *global* fashion (Schütze, 1998; Purandare and Pedersen, 2004). In the *local* variant, for each sub-corpus T of the HILS dataset, the word vectors making up the word-matrix are computed from occurrences in T, using all its non-stop words, $NS(T)$, as the features Σ_{f_2}. In the *global* variant, word vectors are instead computed from the NYT corpus, using $NS(NYT) \cap NS(T)$ for features of the word vectors, where $NS(NYT)$ are the 20k most frequent non-stop words in the NYT corpus[7]. The idea is to have the word vectors determined from a much larger data set, containing word-occurrences that are not constrained to be in the vicinity of one of the words in the HILS dataset. For the pseudoword data, the word vectors were always computed in the local fashion.

WORD	PARALLELISM				SPREAD				
	LOCAL		GLOBAL			LOCAL		GLOBAL	
	$S^1 \parallel S^2$	$S^1 \parallel S^2_{\widehat{W}}$	$S^1 \parallel S^2$	$S^1 \parallel S^2_{\widehat{W}}$	S^1	S^2	$S^2_{\widehat{W}}$	S^2	$S^2_{\widehat{W}}$
hard	.14	.18	.48	.49	.23	.99	.99	.98	.98
interest	.27	.33	.36	.38	.13	.85	.85	.91	.95
line	.35	.33	.40	.42	.14	.93	.95	.95	.96
serve	.50	.45	.50	.54	.14	.74	.85	.89	.87
animal_river	.64	.68			.42	.99	.99		
banana_moon	.45	.49			.23	.99	.98		
data_school	.62	.64			.26	.96	.96		
railway_admission	.57	.59			.28	.98	.98		
rely_illustration	.61	.61			.14	.79	.95		

Table 1: Summary of geometric experiment results

Under parallelism, local in Table 1, column $\mathbf{S^1} \parallel \mathbf{S^2}$ reports the average of cosine measures between first-order representations and second-order representations, $\frac{1}{N}\sum_{i=1}^{N} \cos\left[\mathbf{S^1}(\mathcal{P}_i), \mathbf{S^2}(\mathcal{P}_i)\right]$, for all senses s_i of each target word in the local variant. Column $\mathbf{S^1} \parallel \mathbf{S^2}_{\widehat{\mathbf{W}}}$ shows outcomes when the rows of the word-matrix are L2-normalised (euclidean length). global gives the global variant. For neither the local nor the global variants would one say that these cosine values indicate that the derived $\mathbf{S^1}$ and $\mathbf{S^2}$ vectors are approximately *parallel*. For most of the HILS items, the average cosine scores increased in moving from the local to the global calculation of the word vectors, but still did not result in approximately parallel vectors. For the pseudowords, although the parallelism is higher, they still cannot be fruitfully thought of as approximately parallel to each other.

Spread Another comparison that can be made concerns the angular spread, as measured by cosine similarity, amongst the sense vectors for different senses of a given word. The right-hand side of Figure 3 shows examples of low and high spread of two same-order sense vectors for different senses of the word *hard*. Intuitively, a high angular spread will benefit a sense disambiguation or discrimination algorithm, whereas a low spread will make distinguishing amongst senses more difficult. In this work, spread is measured by taking the average of pairwise cosine measures between sense vectors of the same order that represent different senses of an ambiguous word. Table 1 gives the outcomes, with the same settings used as for the consideration of parallelism. Under spread, local and global the cosine averages are given for each ambiguous item for the sense vector types $\mathbf{S^1}$, $\mathbf{S^2}$ and $\mathbf{S^2_{\widehat{W}}}$. For both the HILS and the pseudoword data, first-order sense representations are far more spread (lowest average cosine scores) than their second-order counterparts, with the global variants being still even less spread than the local.

In order to further contrast these representations, we perform supervised word-sense disam-

[7]using $20k$ to abbreviate 20×10^3

biguation and unsupervised word-sense discrimination experiments. In all of these experiments we employed context vectors and word matrices using the same local and global dimensions used in the geometry experiments above.

3.3. Supervised word-sense disambiguation experiments

A Rocchio classifier was implemented: for each sense in training data, a sense vector is computed ($\mathbf{S^1}$ or $\mathbf{S^2}$), and then the context vector ($\mathbf{C^1}$ or $\mathbf{C^2}$) of an ambiguous test-item is categorised by assignment to the nearest candidate sense vector, as measured by cosine.

Experiments were done both with HILS data and the pseudoword data. The data for an ambiguous item was randomly split into 60% training and 40% test. To ensure robustness, four independent splits were done and results are reported as averages over these splits. Performance is evaluated via a precision score representing the percentage of test context vectors assigned to their correct senses. In computing $\mathbf{C^2}$ vectors, the local approach was taken, so with word vectors computed from the sub-corpus of occurrences of the ambiguous item. The outcomes are shown in Table 2 under the supervised header.

For the HILS and pseudoword data, $\mathbf{C^1}$ vectors outperformed $\mathbf{C^2}$ and $\mathbf{C^2_{\widehat{W}}}$ vectors. And in three out of four HILS cases, and all pseudoword cases, $\mathbf{C^2_{\widehat{W}}}$ vectors outperformed $\mathbf{C^2}$ vectors. These results mirror to a large extent the behaviours observed in the geometric experiments, specifically in the spread experiments: since second-order sense vectors are not very spread out, it is difficult for the Rocchio classifier to correctly assign a sense to an instance.

For the pseudoword data the gap between $\mathbf{C^1}$ and $\mathbf{C^2_{\widehat{W}}}$ was smaller, and overall the pseudoword results are across the board significantly higher than the HILS target word results. This is consistent with previous research that has found that word-sense disambiguation experiments done on pseudowords tend to report higher results than the same experiments done on real ambiguous words (Gaustad, 2001).

A majority sense baseline can also be seen in Table 2 under column **M**. $\mathbf{C^1}$ vectors outperform this baseline more often than the other two context vectors, but $\mathbf{C^2_{\widehat{W}}}$ come at a close second place.

3.4. Unsupervised word-sense discrimination experiments

For the unsupervised experiments a training set of context vectors of an ambiguous item are clustered via the K-Means algorithm[8]. Assuming a 1-to-1 sense/cluster relationship, K is set to the number of senses of the ambiguous item. In the standard K-Means formulation, the metric that decides the assignment of a data point to a cluster is the L2 (euclidean length) distance. However, to ensure symmetry with the supervised experiments we also carried out experiments using cosine as the assignment metric. A clustering is evaluated using Purandare and Pedersen's (2004) method: items from a test set are assigned to their nearest cluster centres and for each possible sense-to-cluster mapping, a precision score on the test set is determined, with the maximum of these reported as the final score. Two types of experiments are done using the HILS data. A 'local' type uses the same training-test splits as in section 3.3; results are reported in Table 2 under (**trn & tst: HILS**). A 'global' type trains on the NYT and then uses the full HILS target word sub-corpora for the evaluation; results are reported under (**trn: NYT, tst: HILS**). Ex-

[8] We used a modified version of Wei Dong's implementation: http://www.cs.princeton.edu/~wdong/kmeans/

WORD	M	SUPERVISED			UNSUPERVISED (L2)						UNSUPERVISED (cos)					
		trn & tst: HILS			trn & tst: HILS			trn: NYT, tst:HILS			trn & tst: HILS			trn: NYT, tst:HILS		
		C^1	C^2	$C^2_{\widehat{W}}$	C^1	C^2	$C^2_{\widehat{W}}$	C^1	C^2	$C^2_{\widehat{W}}$	C^1	C^2	$C^2_{\widehat{W}}$	C^1	C^2	$C^2_{\widehat{W}}$
hard	80	79	76	69	79	60	52	77	62	45	57	66	71	66	62	51
interest	53	78	60	71	46	32	42	50	47	59	47	55	49	48	49	59
line	53	77	48	69	50	31	36	53	39	51	43	38	44	44	47	52
serve	41	81	63	70	39	38	50	44	50	62	54	54	53	59	47	61
AVERAGE	*57*	*79*	*62*	*70*	*54*	*40*	*45*	*56*	*49*	*55*	*50*	*53*	*54*	*54*	*51*	*56*
		trn & tst: NYT			trn & tst: NYT						trn & tst: NYT					
PSEUDOWORD	M	C^1	C^2	$C^2_{\widehat{W}}$	C^1	C^2	$C^2_{\widehat{W}}$				C^1	C^2	$C^2_{\widehat{W}}$			
animal_river	67	85	65	79	67	67	63				67	66	57			
banana_moon	76	88	66	82	75	74	75				69	73	65			
data_school	79	83	82	87	73	77	68				58	77	64			
railway_admission	86	90	68	83	83	71	56				70	77	58			
rely_illustration	78	90	85	87	78	79	81				82	83	79			
AVERAGE	*77*	*87*	*73*	*84*	*75*	*74*	*69*				*69*	*75*	*65*			

Table 2: Performance results for Supervised Rocchio word-sense disambiguation experiments and Unsupervised K-Means word-sense discrimination experiments

periments are also done with the NYT pseudoword data. For increased robustness, we run each K-Means experiment 10 times with different sets of randomly chosen initial cluster centres. The evaluation scores of these 10 runs are averaged. For the local experiments, the 10-run averages for each splitting are in turn averaged again to provide an overall precision score for each target word.

The top part of Table 2 shows L2 and cosine HILS results, under unsupervised (l2) and unsupervised (cos), respectively, which are then divided into local experiments results (**trn & tst: HILS**) and global experiments results (**trn: NYT, tst: HILS**). It is clear from the table that supervised results are far superior to the unsupervised results, with the $\mathbf{C^1}$ roughly following the distribution of the predominant sense (column **M**). In general, it is difficult for context-based WSD systems to outperform this baseline (McCarthy et al., 2004). It seems that global training tends to benefit all three types of context vectors when clustering by L2 K-Means, even if its benefit is modest for cosine K-Means. Within the L2 K-Means case, the local $\mathbf{C^1}$ vectors perform better than the other two types of second-order context vectors, largely reflecting the geometry predictions. However, the global $\mathbf{C^2_{\widehat{W}}}$ turns out somewhat comparable to $\mathbf{C^1}$. Across the cosine K-Means experiments (local and global), results for the two types of second-order context vectors have mixed performance in relation to the baseline but remain comparable to $\mathbf{C^1}$. It can be seen that $\mathbf{C^2_{\widehat{W}}}$ vectors outperform the baseline 7 times, followed by $\mathbf{C^1}$ at 4 times and then the $\mathbf{C^2}$ at 3 times. The target word where the baseline is outperformed more often is *serve*, which is the word that has a more balanced sense distribution (see Figure 2).

In the pseudowords (bottom part of Table 2), $\mathbf{C^2}$ vectors follow the baseline closely in L2 and cosine K-Means. L2 $\mathbf{C^1}$ scores follow the baseline closer than cosine $\mathbf{C^1}$ scores, while $\mathbf{C^2_{\widehat{W}}}$ scores tend to stray below the baseline for both L2 and cosine K-Means, with the notable exception of *rely_illustration*, a pseudoword that tends to have good results across the board possibly because it is made up of words with mixed parts of speech (a verb and a noun), making discrimination easier.

4. Comparisons and Conclusions

Purandare and Pedersen (2004) report a number of experiments on word-sense discrimination, and contrast first-order and second-order outcomes. They evaluate on two separate sense-tagged corpora: a version of the HILS dataset and a smaller dataset derived from the SENSEVAL-2 dataset which has 10 times fewer examples per sense than the HILS dataset. They found that second-order outcomes exceeded first-order on the smaller data set, but that first-order outcomes exceeded second-order on the larger data set. As we argue below, however, inspection of their definitions reveals that they are not really comparing minimal pairs in the first and second order versions.

In all our experiments the features of vectors have been identifiable with *unigrams*. Purandare and Pedersen (2004) work also with features they term *co-occurrences*. Each such feature is identifiable with an unordered pair $\{x, y\}$ and a stretch of text instantiates the feature if x and y occur within a small distance of each other, in any order[9]. Given a chosen set of co-occurrence features, $\mathcal{F}_{co}$, a word-matrix suitable for use in the construction of $\mathbf{C^2}$ vectors can be constructed by building a matrix $\mathbf{WM}_{co}$ such that $\mathbf{WM}_{co}[x][y]$ records the frequency of the co-occurrence feature $\{x, y\}$ in an entire corpus[10]. As we have done, they then represent a 20-word window centred at a particular token with a $\mathbf{C^2}$ vector derived by summing the rows of $\mathbf{WM}_{co}$ for the words in that window. Let $\mathcal{F}_{uni}$ be the unigrams which appear in the chosen set of co-occurrence features $\mathcal{F}_{co}$. Their procedure for constructing $\mathbf{C^2}$ vectors is equivalent to multiplying a $\mathbf{C^1}$ vector using unigram features $\mathcal{F}_{uni}$ by the word matrix $\mathbf{WM}_{co}$ (definition 5):

$$\mathbf{C}^2(p) = \mathbf{C}^1_{\mathcal{F}_{uni}}(p) \times \mathbf{WM}_{co}$$

The natural comparison to make is between outcomes with these $\mathbf{C^2}$ vectors and outcomes with their $\mathbf{C}^1_{\mathcal{F}_{uni}}$ counterparts before post-multiplication by the word-matrix. However, this is not the comparison made by Purandare and Pedersen (2004): in their *first-order* experiments, they represent the 20-word window centred around an occurrence of w by the values of *co-occurrence features* $\{x, y\}$ in that window[11]. These are $\mathbf{C}^1_{\mathcal{F}_{co}}$ vectors.

Thus it seems fair to say that the conclusions they draw concerning dependency on the size of the data set are not based on contrasting minimal pairs. Rather than contrasting outcomes with a particular kind of $\mathbf{C^1}$ vector with those that would be obtained simply by post-multiplying *exactly those* $\mathbf{C^1}$ *vectors* by some kind of word-matrix, they are contrasting co-occurrence based $\mathbf{C}^1_{\mathcal{F}_{co}}$ with the post-multiplication of unigram-based $\mathbf{C}^1_{\mathcal{F}_{uni}}$. Their findings concerning dependency on the size of the data-set are thus potentially attributable to factors other than the first-order vs. second-order contrast. In our experiments we did not systematically vary the size of the data set and it remains for future work to revisit this size dependency issue with more strictly comparable first- and second-order representations.

In this work, we contrasted $\mathbf{C^1}_{\mathcal{F}_{uni}}$ versus $\mathbf{C^2}_{\mathcal{F}_{uni}}$. The geometric experiments as well as the supervised word-sense disambiguation experiments suggest that in this simplest configuration,

[9] Mostly this distance is set to be 3 words or less. They also work with an ordered variant, of which detailed discussion we omit for space reasons.

[10] Instead of a count, a statistical association score might be recorded. They also further apply SVD-based dimensionality reduction to obtain $\widehat{WM}$, with a reduced dimensionality version of each row of WM. For the point we wish to make these details do not matter.

[11] The x and y of the co-occurrence feature are unrelated to the word w which centres the context, save for needing to be instantiated in the window around that occurrence of w

$\mathbf{C}^1$ vectors are better than $\mathbf{C}^2$ vectors. In the *supervised* WSD experiments, the $\mathbf{C}^1$ vectors beat both variants of $\mathbf{C}^2$ vectors on all of the HILS words and on 4 of the pseudowords, out of a total of 5. On both datasets $\mathbf{C}^1$ vectors beat the M baseline in 8 out of 9 cases and $\mathbf{C}^2_{\widehat{\mathbf{W}}}$ does so in 7 out of 9 cases.

In the *unsupervised* experiments, on the average the $\mathbf{C}^1$ vectors and $\mathbf{C}^2$ vectors outcomes are much closer together, and for the HILS data, best outcomes are about 25% down from the supervised case, whilst for the pseudoword data, the fall is around 12%; thus the multiway ambiguity of the HILS data versus the 2-way ambiguity of the pseudoword data seems to be particularly challenging to the unsupervised methods. On the pseudowords, the M baseline is seldom beaten, and the advantage of $\mathbf{C}^1$ vectors over $\mathbf{C}^2$ vectors from the supervised case is not replicated. On the HILS data, only for *serve* is the M baseline often beaten. Across the representations, the 'global' version with clustering on the large NYT corpus performed better than the 'local' version, which clusters on a subset of the HILS. And again, the advantage of $\mathbf{C}^1$ vectors over $\mathbf{C}^2$ vectors from the supervised case is not replicated, with varying outcomes across the words, and a close final average.

Thus these experiments have shown an advantage of $\mathbf{C}^1$ vectors over $\mathbf{C}^2$ vectors for the supervised case, and no clear winner for the unsupervised case. It has to be stressed that the setting used for $\mathbf{C}^1$ vectors and $\mathbf{C}^2$ vectors were in many respects, the simplest possible, and a different picture might emerge under different settings. In future work we will embark on experiments exploring those settings, such as weighting schemes, alternative feature selection schemes, the effects of applying SVD to the different objects involved and alternative $\mathbf{C}^2$-construction operations to summation and averaging.

References

de Marneffe, M.-C. and Dupont, P. (2004). Comparative study of statistical word sense discrimination. In *Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data (JADT 2004)*.

Gaustad, T. (2001). Statistical corpus-based word sense disambiguation: Pseudowords vs. real ambiguous words. In *Companion Volume to the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001) – Proceedings of the Student Research Workshop*, Toulouse.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martinez, D. and Baldwin, T. (2011). Word sense disambiguation for event trigger word detection in biomedicine. *BMC Bioinformatics*, 12 Suppl 2.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona.

Oh, J.-H. and Choi, K.-S. (2002). Word sense disambiguation using static and dynamic sense vectors. In *COLING*.

Purandare, A. and Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of CoNLL-2004*, pages 41–48. Boston, MA, USA.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322.

Sagi, E., Kaufmann, S., and Clark, B. (2008). Tracing semantic change with latent semantic analysis. In *Proceedings of ICEHL 2008*.

Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, Upper Saddle River, NJ.

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

Sugiyama, K. and Okumura, M. (2009). Semi-supervised clustering for word instances and its effect on word sense disambiguation. In *CICLing*, pages 266–279.

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

Wang, T. and Hirst, G. (2010). Near-synonym lexical choice in latent semantic space. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING ’10, pages 1182–1190, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, pages 266–271, Plainsboro, NJ. Morgan Kaufmann Publishers.