Expectation Vectors: A Semiotics Inspired Approach to Geometric Lexical-Semantic Representation

Justin Washtell
University of Leeds
Leeds, UK
washtell@comp.leeds.ac.uk

Abstract

We introduce a new family of geometric models of meaning, inspired by principles from semiotics and information theory, based on what we call Expectation Vectors. We present theoretical arguments in support of these representations over traditional context-feature vectors: primarily that they provide a more intuitive representation of meaning, and detach vector representation from the specific context features thereby allowing arbitrarily sophisticated language models to be leveraged. We present a preliminary evaluation of an expectation vector based word sense disambiguation system using the SemEval-2007 task 2 dataset, with very encouraging results, particularly with respect to ambiguous verbs.

1 Introduction

It is a cornerstone assumption of distributional lexical semantics that the distribution of words in a corpus reflects their meaning. Common interpretations of this include the Distributional Hypothesis (Harris, 1954) and the Contextual Hypotheses (Miller & Charles, 1991), which state that there is a relationship between a word's meaning, and the context(s) in which it appears. In recent years this insight has been borne out by correlations between human judgements and distributional models of word similarity (Rapp, 2002), and steady advances in tasks such as word sense disambiguation (Schütze, 1998) and information retrieval. The workhorse of these approaches are wordspace models: vectors built from context features which serve as geometric analogues of meaning. Despite many advances, substantial problems exist with this approach to modelling meaning. Amongst these are the problems of data sparseness and of how to model compositional meaning.

In this short paper, we introduce a new family of wordspace models, based on insights gleaned from semiotics and information theory, called Expectation Vectors. These retain the convenient vector-based paradigm whilst encouraging the exploitation of advances in language modelling from other areas of NLP. We finish by outlining some present efforts to evaluate expectation vectors in the area of word sense disambiguation.

2 Modelling meaning from context

Perhaps one of the most prominent application areas to exploit context-based wordspace models is that of word sense induction and disambiguation (WSI/WSD). The prevailing approach to this problem is based on a fairly literal interpretation of the Distributional Hypothesis: that is to cluster or classify instances of ambiguous words according to certain features of the context in which they appear – invariably other words. It is not difficult to see why this approach is limiting: as Pedersen (2008) observes, *"the unifying thread that binds together many short context applications and methods is the fact that similarity decisions must be made between contexts that share few (if any) words in common."* This is a manifestation of what is commonly referred to at the data sparseness problem, and it pervades all of corpus-based NLP. This problem is exacerbated as available examples of a word sense decrease, or finer sense granularities are sought. For supervised tasks this implies that a large training set is required, which is often expensive. For unsupervised tasks, such as WSI, it has negative implications for cluster quality and rule learning. Consequently, Leacock *et al* (1996) observe that WSD systems which operate directly upon context are: *"plagued with the same problem, excellent precision but low recall"*.

"Backing off" to more general feature classes through say lemmatization or part-of-speech tagging affords one way of alleviating sparseness (Joshi & Penstein-Rosé, 2009), assuming these features are pertinent to the task. Similar strategies include the use of dual-context models where immediate lexical features are backed up by more general topical ones garnered from the wider context of the ambiguous word (Leacock *et al*, 1996; Yarowsky, 1993).

Others have tackled the problem of sparseness without recourse to generalized feature classes,

Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, ACL 2010, pages 45–50,
Uppsala, Sweden, 16 July 2010.

through the exploitation of higher-order distributional information. Schütze (1998) popularised this approach within the WSD/WSI task. Rather than comparing contexts directly, it is the distributional similarity of those features (in the corpus) which are compared. Specifically, Schütze composed context vectors by summing the vectors for every word in a context, where those vectors were themselves formed from the total of word co-occurrence counts pertaining to every instance of that word in the corpus. The resultant context vectors are therefore comparatively dense, and carry second-order information which makes otherwise unlike contexts more amenable to comparison. One contention of this model is that it conflates co-occurrence information from all occurrences of a word in the corpus, regardless of their sense. The defence is that because the actual senses of the term instances which appear in the context of the ambiguous word will tend to be pertinent to that word's own specific sense, it is that common aspect of their respective conflated-sense vectors - when summed - which will dominant the resultant context vector. Purandare & Pedersen (2001) performed a comparative study of disambiguation approaches based on first-order context, and on second order context as per Schütze (1998). They found that while Schütze's approach provided gains when data was limited, when the training corpus was large enough that sufficient examples existed, clustering on first order context was actually a better approach. This suggests that while alleviating the data-sparseness problem, the practice of expanding context vectors in this way introduces a certain amount of noise, presumably by inappropriately over-smoothing the data.

Another approach to the sparse data problem which was also part of Schütze's framework is dimensionality reduction by Singular Value Decomposition (SVD). In SVD the set of context features are analytically combined and reduced in a manner that exploits their latent similarities, whereafter traditional vector measures can be used. Very similar techniques to both of those used by Schütze have been used for query expansion and document representation in information retrieval (Qiu & Frei, 1993; Xu *et al*, 2007).

Several variations upon Schütze's approach to WSD have been explored. Dagan *et al* (1995) and Karov & Edelman (1996) both apply what they call "similarity-based" methods which, while markedly different on the surface to that of Schütze, are similar in spirit and intent. Karov & Edelman, for example, use machine-readable dictionary glosses as opposed to corpus-derived co-occurrences, and apply an iterative bootstrapping approach to augment the available data, rather than strict second-order information.

Typically, context vectors comprise a component (dimension) for each designated feature in a word's context. In a simple bag-of-words model this might equate to one vector component for each potential word that can appear in the context. For more sophisticated n-gram or dependency-based models, which attempt to better capture the structure inherent in the language, this number of vector components must be increased. The more sophisticated the language model becomes therefore, the more acute the sparse data problem. Techniques like SVD can reduce this sparseness, but other issues remain. How does one weight heterogeneous features when forming a vector? How does one interpret vectors reduced by SVD? Looking at the variety of approaches to tackling the problem, we might be forgiven for questioning whether representing meaning as a vector of context features is in fact an ideal starting point for semantic tasks such as WSD.

In the following section we describe a means of entirely detaching context feature selection from vector representation, such that an arbitrarily sophisticated language model can be used to generate dense, comparable vectors. Necessarily, we also present a prototype distributional language model that will serve as the basis of our investigations into this approach.

3 System & approach

3.1 Lexical Expectation Vectors

Theoretical motivation. The motivation behind the method presented herein comes both from the fields of semiotics and information theory. It is the notion that the "meaning" of an utterance is not in the utterance itself, nor in its individual or typical context; it is in the *disparity* between our expectations based on that context, and the utterance (Noth, 1990; Chandler, 2002). Meaning in this sense can be seen as related to information (Attneave, 1959; Shannon, 1948): an utterance which is entirely expected under a regime where speaker and interpreter have identical frames of reference communicates nothing; conversely an extremely creative utterance is laden with information, and may have multiple non-obvious interpretations (poetry being a case in point - Riffaterre, 1978). This idea is also lent some weight by psycholinguistic experiments which have revealed correlations between a word's disparity from its preceding context, and processing times in human subjects. Similar insights have been employed in some very recent

attempts to model compositional word meaning Erk & Padó (2008) and Thater *et al* (2009). These models augment word and context representations with additional vectors encoding the selectional preferences (expectations) pertaining to the specific syntactic/semantic roles of the participating words. So far these systems rely upon parsed corpora and have been tested only with very limited contexts (e.g. pairs of words having specific dependency relations).

Lexical expectation vectors are based on a similar and very simple premise: rather than building a vector for a context by conflating the features which comprise the various context words (as per Schutze, 1998), we instead conflate all the words which might be expected to appear *within the context* (i.e. in the headword position). Consider the following short context taken from the SemEval-2007 task 2 dataset:

```
Mr. Meador takes responsibility
for <?> and property management .
```

The strongest twenty elements of its expectation vector (as generated by the system described below) are shown in table 1. The figures represent some measure of confidence that a given word will be found in the headword position <?>.

0.42	education	0.31	chancellor
0.38	forms	0.31	routine
0.36	housing	0.31	health
0.35	counselling	0.31	research
0.35	these	0.31	assessment
0.35	herself	0.3	detailed
0.34	database	0.3	management
0.33	injuries	0.3	many
0.32	advice	0.3	training
0.31	this	0.3	what

Table 1: An example of an expectation vector.

We make the supposition that when the vectors implied by the respective likelihoods of *all words* implied by two contexts are identical, the contexts can be considered semantically equivalent.[1] Note that the actual headword appearing in the context is not taken into consideration for the purposes of calculating expectation. In this example it occur at rank 62 out of ~650,000, implying that its use in this context is not atypical.

Formal approach. For the purpose of our present research, we adopt the following formal framework for generating an expectation vector. Given a context c, each component of the expectation vector $\mathbf{e}$ arising from that context is estimated thusly:

$$\mathbf{e}_j = P(j|c) \sim \max_{o_i^k \in O_i} \text{sim}(o_j^k, c)$$

Where j is a given word type in the lexicon, O_j is the set of all observed contexts of that word type in some corpus, o_j^k is the k^{th} observed context of that word type, and sim(o,c) is some similarity measure between two contexts.

The process of generating an expectation vector can be thought of as a kind of transform from *syntagmatic* space, into *paradigmatic* space. This mapping need not be trivial: items which are close in the syntagmatic space need not be close in the paradigmatic space and vice-versa (although in practice we expect some considerable correlation by virtue of the distributional hypothesis). Note that although our work herein assumes a popular vector representation of context, the nature of the contexts and the similarity measure which operates upon them are not constrained in any way by the framework given above. For example they may equally well be dependency trees.

In the following section we outline a distance-based language model comprising a context model and a similarity metric which operates upon it. This choice of model allows us to maintain a purely distributional approach without suffering the data-sparseness associated with n-gram models.

3.2 Language model

Theoretical motivation. The precise relationship between syntagmatic and paradigmatic spaces implied by the expectation transform depends upon the language model employed. In a naive language model which assumes independence between features, this mapping can be fully represented by a square matrix over word types. Although such models are the mainstay of many systems in NLP, adopting the toolset of an expection transform in such a case gains us little. Therefore the relevance of the approach to the present task depends wholly upon having a suitably sophisticated language model.

Building on the work of Washtell (2009) and Terra & Clarke (2004), a distance-based language model is used in the present work. This is in contrast to the bag-of-words, n-gram, or syntactic dependency models more commonly described in the NLP literature. There are two hypothesised advantages to this approach. Firstly, this avoids the issue of immediate context versus wider topical

[1] Equivalent with respect to the head of the context. This is not the same as saying the passages have the same meaning, which requires recourse to compositionality.

context. While immediate context is generally accepted to play a dominant role in WSD, both near and far context have been shown to be useful - the specific balance being somewhat dependent on the ambiguous word in question (Yarowsky, 1993; Gale et al, 1992; Leacock *et al*, 1996). As Ide & Veronis (1998) astutely observe, "*although a distinction is made between micro-context and topical context in current WSD work, it is not clear that this distinction is meaningful. It may be more useful to regard the two as lying along a continuum, and to consider the role and importance of contextual information as a function of distance from the target.*" This is precisely the assumption adopted herein. Secondly, the use of distance-based information alleviates data sparseness. This is simply by virtue of the fact that all words types in a document form part of a token's context (barring document boundaries, no cut-off distance is imposed). Moreover, as it is specific distance information which is being recorded, rather than (usually low) frequency counts, context vector components and the similarity measurements which arise from them exhibit good precision. Washtell (2009) showed that these properties of distance-based metrics lead to measurable gains in information extracted from a corpus. In the context of modelling human notions of association this also led to improved predictive power (Washtell & Markert, 2009).

Formal approach. We do not pre-compute any statistical representation of the data upon which our language model draws. With available approaches this would either require throwing away a large number of potentially relevant higher-order dependencies, or would otherwise be intractable. Our intuition is that the truest representation of the language encoded in the corpus is the corpus itself. We therefore use an indexed corpus directly for all queries.

We use the following as a prototype measure of structural similarity (see section 3.1), although note that others are by all means possible.

$$\mathrm{sim}(\mathbf{o},\mathbf{c}) = \frac{\sum_{\{p,q\}\subseteq O\cap C} f(\mathbf{o}_p, \mathbf{o}_q, \mathbf{c}_p, \mathbf{c}_q)}{\min(|O|, |C|)^2}$$

Where **o** and **c** are context vectors whose *j* components each specify the position in the text of the nearest occurrence (to the head of the context) of a given word type. *O* and *C* are the set of indices of all non-zero (i.e. observed) components in **o** and **c** respectively. The head of the context is represented by an additional component in vectors **o** and **c**, and is always treated as observed. *f* is a further function of the positions of words *p* and *q* in both contexts. It returns a similarity score in the unit range designating how similar the distance $\mathbf{o}_p \leftrightarrow \mathbf{o}_q$ is to that of $\mathbf{c}_p \leftrightarrow \mathbf{c}_q$.

The more consistent the relative positions of the various symbols comprising two contexts, the stronger their similarity. Note that the measure is additive: symbols which occur at all in both contexts result in positive score contributions. We assume that a context is usually incomplete (i.e. that that which lies outside it is unknown, rather than non-existent). The minimum operator in the denominator (the normalization factor) therefore ensures that words present only in the larger of two contexts do not constitute negative evidence.

This formulation allows for considerable leeway in how word distances are represented and compared. In this work we choose to treat distances proportionately, so small variations in word position between distant (presumably topically related words) are tolerated better than similar distance variations between neighbouring (more syntactico-semantically related) words.

4 Word Sense Disambiguation

A WSD system based on expectation vectors was ineligible in the SemEval-2010 WSI/WSD task by virtue of restrictions disallowing the use of a corpus-based language model. Instead, this task implicitly encouraged participants to focus on context feature selection and clustering approaches. It seems unlikely to us that these stages are where the major bottlenecks for WSD (or WSI) lie; performing WSD on short contexts without any extra-contextual information (i.e. general linguistic or domain experience) is arguably not a task which even humans could be expected to perform well. For this reason we have chosen to focus initially on the well explored SemEval-2007 task 2 dataset.

4.1 Preliminary Evaluation

An expectation vector was produced for each training and test instance in the SemEval dataset by matching the headword's context against that of each word position in the British National Corpus using an implementation of the distance based similarity measure outlined in section 3.2. For matters of convenience, independent forwards and backwards expectation vectors were produced from the context preceding the headword and that following it, and their elements were multiplied together to produce the final vector. No lemmatization or part-of-speech tagging was

employed. Neither was any dimensionality reduction, each vector therefore having ~650,000 elements: one for each word type in the corpus.

Each test sample's vector was compared against all corresponding training sample vectors using both cosine similarity and Euclidean distance[2]. In the MAX setups (see Table 2), each test case was assigned the sense of the single nearest training example according to the metric being used. In the CosOR setup, sense scores were generated by applying a probabilistic OR operation over the squared Cosine similarities of *all* relevant training examples[3]. The BaseMFS setup is a popular baseline in which the most frequent sense in the training set for a given ambiguous word is attributed to every test case.

	Nouns	Verbs	All
CosMAX	**83.6** ▲6.1 ▼22.8	**70.5** ▲7.6 ▼14.4	**79.5** ▲6.7 ▼19.5
EucMAX	78.9	67.0	75.1
CosOR	83.5	66.1	78.0
BaseMFS	78.8	65.5	74.5

Table 2: Recall on SemEval WSD task, including relative performance gain (▲) and error reduction (▼) over baseline for best setup (preliminary based on first 25% of test cases).

	Nouns	Verbs	All
BEST	**86.8** ▲7.3 ▼30.9	**76.2** ▲0.0 ▼0.0	**81.6** ▲3.7 ▼13.6
BaseMFS	80.9	76.2	78.7

Table 3: Recall of best official SemEval WSD systems (Agirre & Soroa, 2007), showing relative performance gain and error reduction over baseline.

Table 2 shows the results for each test case in terms of recall, for all words and for nouns and verbs separately. Also shown in table 3 are the best and baseline figures for the official entries from the Semeval workshop. Note that figures are not directly comparable between tables because our preliminary results represent only the first 25% of the SemEval dataset (hence the different baselines). To aid some comparison, figures are included in both tables indicating the relative increases in recall over the baseline, and relative reduction in error. Note that the system employed here is not a word sense induction system as were most of those participating in the official SemEval task. The setup of the tasks however allows for systems which perform poorly under the induction evaluation to perform competitively as disambiguation systems, so we are not precluded from making meaningful comparisons here.

5 Discussion and Future Direction

We have presented a new type of wordspace model based on vectors derived from the predictions of a language model applied to a context, rather than directly from the features of a context itself. We have conducted a preliminary investigation of the semantic modelling power of such vectors in the setting of a popular WSD task. The results are very encouraging. Although it is too early to draw hard conclusions, preliminary results suggest a performance at least comparable the present state of the art on this task. What is particularly noteworthy is that the approach taken here seems to perform equally well at discriminating verbs and nouns. Verbs have traditionally proven very problematic: *none* of the six SemEval systems were able to improve upon the verb baseline. More recent studies have focused on discriminating nouns (Brody & Lapata, 2009; Klapaftis & Manandhar, 2007).

Further gains might be expected by employing a corpus which is more closely matched to the material being disambiguated, such as the Wall Street Journal in the present case.

It is also worth noting that the system presented here was aided only by an untagged un-lemmatised corpus, without the use of any structured knowledge sources. While we expect that judicious use of lemmatization could improve these results, we believe the key to the quality of expectation vectors is in the specific predictive language model employed. We have scarcely experimented with this, opting for a relatively untested distance-based model throughout, and choosing instead to experiment with the application of different vector similarity measures. While the nature of the language model used enables it to capture complex interdependencies, and long-range dependencies, it is based on direct querying of a corpus and therefore does not scale at all well. This makes its use in the context of most applications or with larger corpora untenable. Exploring alternative language models (drawing upon the copious research in this field) is therefore a focus for future research; the ability to do this highlights one of the major advantages of this approach to modelling meaning.

[2] Cosine Similarity captures the similarity between the relative proportions of features present in each of two vectors. By contrast, Euclidean Distance compares the actual values of corresponding features.

[3] Although encountered rarely in the literature, squared Cosine Similarity is a pertinent quantity for tasks that go beyond simple ranking. As with Pearson's R^2, it represents the degree or proportion of similarity (consider that the square of an angle's cosine and that of its sine total 1).

References

Eneko Agirre, Airotr Soroa, 2007, *SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems*

Fred Attneave, *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results,* Holt, New York

Samuel Brody, Mirella Lapata, 2009, *Bayesian Word Sense Induction,* Proceedings of EACL 2009, pages 103-111

Daniel Chandler, 2002, *Semiotics: The Basics*, p88, Routleadge

Ido Dagan, Shaul Marcus, Shaul Markovitch, 1995, *Contextual Word Similarity and Estimation from Sparse Data,* Proceedings of 31st ACL, pages 164-171

Katrin Erk, Sebastian Padó, 2008, *A Structured Vector Space Model for Word Meaning in Context*, Proceedings on EMNLP 2008

William Gale, Kenneth Church, , David Yarowsky, *1992, A Method for Disambiguating Word Senses in a Large Corpus, Computers and the Humanities,* 26, 415-429

Zellig Harris, 1954, *Distributional structure*. Word, 10(23), pages 146-162

Nancy Ide, Jean Veronis, 1998, *Word Sense Disambiguation: The State of The Art,*

Mahesh Joshi, Carolyn Penstein-Rosé, 2009, *Generalizing Dependency Features for Opinion Mining*, Proceeding of the ACL-IJCNLP 2009 Conference Short Papers, pages 313-316

Yael Karov, Shimon Edelman, 1996, *Learning Similarity-Based Word Sense Disambiguation from Sparse Data*

Ioannis Klapaftis, Suresh Manandhar, 2008, *Word Sense Induction using Graphs of Collocations,* Proceedings of ECAI 2008, pages 298-302

Claudia Leacock, Geoffrey Towell, Ellen M. Voorhees, 1996, *Towards Building Contextual Representations of Word Senses Using Statistical Models.* In B. Boguraev and H.Pustejovsky (eds) Corpus Processing for Lexical Acquisition. MIT Press, pages 97-113

G. A. Miller, W. G Charles, 1991. *Contextual correlates of semantic similarity.* Language and Cognitive Processes, *6*, 1-28.

Winfried Noth, 1990, *Handbook of Semiotics*, Indiana University Press, p 142

Reinhard Rapp. 2002. *The computation of word associations: comparing syntagmatic and paradigmatic approaches.* In Proceedings of the 19th international Conference on Computational Linguistics.

Hinrich Schütze, 1998, *Automatic Word Sense Discrimination*, Computational Linguistics, 24(1), pages 97-123

Ted Pedersen, 2008, *Computational Approaches to Measuring the Similarity of Short Contexts: A Review of Applications and Methods,* to appear

Amruta Purandare, Ted Pederson, 2004, *Word sense discrimination by clustering contexts in vector and similarity spaces.* Proceeding of the Conference on Computational Natural Language Learning, pages 41-48

Yonggang Qiu, Hans-Peter Frei, 1993, Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval table of contents, 160-169

Miachael Riffaterre, 1978, *Semiotics Of Poetry*, Methuen

Hae Jong Seo, Peyman Milanfar, 2009, *Training-free, Generic Object Detection using Locally Adaptive Regression Kernels*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 99

Claude E Shannon, 1948, *A Mathematical Theory of Communication*, Bell System Technical Journal, vol. 27, pages 379-423, 623-656,

Egidio Terra, Charles L. A. Clarke, 2004, *Fast computation of lexical affinity models ,* Proceedings of the 20th international conference on Computational Linguistics

Stefan Thater, Georgiana Dinu, Manfred Pinkal, 2009, *Ranking Paraphrases in Context,* Proceedings of the 2009 Workshop on Applied Textual Inference, pages 44-47

Justin Washtell. 2009. *Co-dispersion: A windowless approach to lexical association.* In Proceedings of EACL-2009.

Justin Washtell, Katja Markert. 2009. *Comparing windowless and window-based computational association measures as predictors of syntagmatic human associations.* In Proceedings of EMNLP-2009, pages 628-637.

Xuheng Xu, Xiaodan Zhang, Xiaohua Hu, 2007, *Using Two-Stage Concept-Based Singular Value Decomposition Technique as a Query Expansion Strategy,* AINAW'07

David Yarowsky, 1993, *One Sense Per Collocation,* Proceedings on the Workshop on Human Language Technology, pages 266-271