Context-theoretic Semantics for Natural Language: an Overview

Daoud Clarke
University of Sussex
Falmer, Brighton, UK
daoud.clarke@gmail.com

Abstract

We present the context-theoretic framework, which provides a set of rules for the nature of composition of meaning based on the philosophy of *meaning as context*. Principally, in the framework the composition of the meaning of words can be represented as multiplication of their representative vectors, where multiplication is distributive with respect to the vector space.

We discuss the applicability of the framework to a range of techniques in natural language processing, including subsequence matching, the lexical entailment model of Dagan et al. (2005), vector-based representations of taxonomies, statistical parsing and the representation of uncertainty in logical semantics.

1 Introduction

Techniques such as latent semantic analysis (Deerwester et al., 1990) and its variants have been very successful in representing the meanings of words as vectors, yet there is currently no theory of natural language semantics that explains how we should compose these representations: what should the representation of a phrase be, given the representation of the words in the phrase? In this paper we present such a theory, which is based on the philosophy of *meaning as context*, as epitomised by the famous sayings of Wittgenstein (1953), “Meaning just *is* use” and Firth (1957), “You shall know a word by the company it keeps”. For the sake of brevity we shall present only a summary of our research, which is described in full in (Clarke, 2007), and we give a simplified version of the framework, which nevertheless suffices for the examples which follow.

We believe that the development of theories that can take vector representations of meaning beyond the word level, to the phrasal and sentence levels and beyond are essential for vector based semantics to truly compete with logical semantics, both in their academic standing and in application to real problems in natural language processing. Moreover the time is ripe for such a theory: never has there been such an abundance of immediately available textual data (in the form of the world-wide web) or cheap computing power to enable vector-based representations of meaning to be obtained. The need to organise and understand the new abundance of data makes these techniques all the more attractive since meanings are determined automatically and are thus more robust in comparison to hand-built representations of meaning. A guiding theory of vector based semantics would undoubtedly be invaluable in the application of these representations to problems in natural language processing.

The context-theoretic framework does not provide a formula for how to compose meaning; rather it provides mathematical guidelines for theories of meaning. It describes the nature of the vector space in which meanings live, gives some restrictions on how meanings compose, and provides us with a measure of the degree of entailment between strings for any implementation of the framework.

The remainder of the paper is structured as follows: in Section 2 we present the framework; in Section 3 we present applications of the framework:

- We describe subsequence matching (Section 3.1) and the lexical entailment model of (Dagan et al., 2005) (Section 3.2), both of which have been applied to the task of recognising textual entailment.
- We show how a vector based representation of a taxonomy incorporating probabilistic information about word meanings can be con-

Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, pages 112–119,
Athens, Greece, 31 March 2009. ©2009 Association for Computational Linguistics

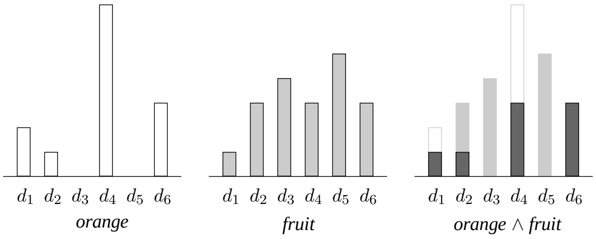


Figure 1: Vector representations of two terms in a space $L^1(S)$ where $S = \{d_1, d_2, d_3, d_4, d_5, d_6\}$ and their vector lattice meet (the darker shaded area).

structed in Section 3.3.

- We show how syntax can be represented within the framework in Section 3.4.
- We summarise our approach to representing uncertainty in logical semantics in Section 3.5.

2 Context-theoretic Framework

The context-theoretic framework is based on the idea that the vector representation of the meaning of a word is derived from the contexts in which it occurs. However it extends this idea to strings of any length: we assume there is some set S containing all the possible contexts associated with any string. A *context theory* is an implementation of the context-theoretic framework; a key requirement for a context theory is a mapping from strings to vectors formed from the set of contexts.

In vector based techniques, the set of contexts may be the set of possible dependency relations between words, or the set of documents in which strings may occur; in context-theoretic semantics however, the set of "contexts" can be any set. We continue to refer to it as a set of contexts since the intuition and philosophy which forms the basis for the framework derives from this idea; in practice the set may even consist of logical sentences describing the meanings of strings in model-theoretic terms.

An important aspect of vector-based techniques is measuring the frequency of occurrence of strings in each context. We model this in a general way as follows: let A be a set consisting of the words of the language under consideration. The first requirement of a context theory is a mapping $x \mapsto \hat{x}$ from a string $x \in A^*$ to a vector $\hat{x} \in L^1(S)^+$, where $L^1(S)$ means the set of all functions from S to the real numbers $\mathbb{R}$ which are finite under the L^1 norm,

$$\|u\|_1 = \sum_{s \in S} |u(s)|$$

and $L^1(S)^+$ restricts this to functions to the non-negative real numbers, $\mathbb{R}^+$; these functions are called the positive elements of the vector space $L^1(S)$. The requirement that the L^1 norm is finite, and that the map is only to positive elements reflects the fact that the vectors are intended to represent an estimate of relative frequency distributions of the strings over the contexts, since a frequency distribution will always satisfy these requirements. Note also that the l_1 norm of the context vector of a string is simply the sum of all its components and is thus proportional to its probability.

The set of functions $L^1(S)$ is a vector space under the point-wise operations:

$$\begin{aligned}(\alpha u)(s) &= \alpha u(s)\\ (u+v)(s) &= u(s) + v(s)\end{aligned}$$

for $u, v \in L^1(S)$ and $\alpha \in \mathbb{R}$, but it is also a lattice under the operations

$$\begin{aligned}(u \wedge v)(s) &= \min(u(s), v(s))\\ (u \vee v)(s) &= \max(u(s), v(s)).\end{aligned}$$

In fact it is a *vector lattice* or *Riesz space* (Aliprantis and Burkinshaw, 1985) since it satisfies the following relationships

$$\begin{aligned}&\text{if} \quad u \leq v \quad \text{then} \quad \alpha u \leq \alpha v\\ &\text{if} \quad u \leq v \quad \text{then} \quad u + w \leq v + w,\end{aligned}$$

where $\alpha \in \mathbb{R}^+$ and $\leq$ is the partial ordering associated with the lattice operations, defined by $u \leq v$ if $u \wedge v = u$.

Together with the l_1 norm, the vector lattice defines an *Abstract Lebesgue space* (Abramovich and Aliprantis, 2002) a vector space incorporating all the properties of a measure space, and thus can also be thought of as defining a probability space, where $\vee$ and $\wedge$ correspond to the union and intersection of events in the σ algebra, and the norm corresponds to the (un-normalised) probability.

2.1 Distributional Generality

The vector lattice nature of the space under consideration is important in the context-theoretic framework since it is used to define a degree of entailment between strings. Our notion of entailment is based on the concept of *distributional generality* (Weeds et al., 2004), a generalisation of the distributional hypothesis of Harris (1985), in which it is assumed that terms with a more general meaning will occur in a wider array of contexts, an idea later developed by Geffet and Dagan (2005). Weeds et al. (2004) also found that frequency played a large role in determining the direction of entailment, with the more general term often occurring more frequently. The partial ordering of the vector lattice encapsulates these properties since $\hat{x} \leq \hat{y}$ if and only if y occurs more frequently in all the contexts in which x occurs.

This partial ordering is a strict relationship, however, that is unlikely to exist between any two given vectors. Because of this, we define a *degree of entailment*

$$\mathrm{Ent}(u, v) = \frac{\|u \wedge v\|_1}{\|u\|_1}.$$

This value has the properties of a conditional probability; in the case of $u = \hat{x}$ and $v = \hat{y}$ it is a measure of the degree to which the contexts string x occurs in are shared by the contexts string y occurs in.

2.2 Multiplication

The map from strings to vectors already tells us everything we need to know about the composition of words: given two words x and y, we have their individual context vectors $\hat{x}$ and $\hat{y}$, and the meaning of the string xy is represented by the vector $\widehat{xy}$. The question we address is what relationship should be imposed between the representation of the meanings of individual words $\hat{x}$ and $\hat{y}$ and the meaning of their composition $\widehat{xy}$. As it stands, we have little guidance on what maps from strings to context vectors are appropriate.

The first restriction we propose is that vector representations of meanings should be composable *in their own right*, without consideration of what words they originated from. In fact we place a strong requirement on the nature of multiplication on elements: we require that the multiplication $\cdot$ on the vector space defines a *lattice-ordered algebra*. This means that multiplication is associative, distributive with respect to addition, and satisfies $u \cdot v \geq 0$ if $u \geq 0$ and $v \geq 0$, i.e. the product of positive elements is also positive.

We argue that composition of context vectors needs to be compatible with concatenation of words, i.e.

$$\hat{x} \cdot \hat{y} = \widehat{xy},$$

i.e. the map from strings to context vectors defines a semigroup homomorphism. Then the requirement that multiplication is associative can be seen to be a natural one since the homomorphism enforces this requirement for context vectors. Similarly since all context vectors are positive their product in the algebra must also be positive, thus it is natural to extend this to all elements of the algebra. The requirement for distributivity is justified by our own model of meaning as context in text corpora, described in full elsewhere.

2.3 Context Theory

The above requirements give us all we need to define a context theory.

Definition 1 (Context theory). *$\langle A, S, \hat{}\,, \cdot\, \rangle$ defines a context theory if $L^1(S)$ is a lattice-ordered algebra under the multiplication defined by $\cdot$ and $\hat{}$ defines a semigroup homomorphism $x \mapsto \hat{x}$ from A^* to $L^1(S)^+$.*

3 Context Theories for Natural Language

In this section we describe applications of the context-theoretic framework to applications in computational linguistics and natural language processing. We shall commonly use a construction in which there is a binary operation $\circ$ on S that makes it a semigroup. In this case $L^1(S)$ is a lattice-ordered algebra with convolution as multiplication:

$$(u \cdot v)(r) = \sum_{s \circ t = r} u(s)v(t)$$

for $r, s, t \in S$ and $u, v \in L^1(S)$. We denote the unit basis element associated with an element $x \in S$ by e_x, that is $e_x(y) = 1$ if and only if $y = x$, otherwise $e_x(y) = 0$.

3.1 Subsequence Matching

A string $x \in A^*$ is called a "subsequence" of $y \in A^*$ if each element of x occurs in y in the same order, but with the possibility of other elements occurring in between, so for example $abba$ is a subsequence of $acabcba$ in $\{a, b, c\}^*$. We denote the set of subsequences of x (including the empty string) by $\mathrm{Sub}(x)$. Subsequence matching compares the subsequences of two strings: the more subsequences they have in common the more similar they are assumed to be. This idea has been used successfully in text classification (Lodhi et al., 2002) and recognising textual entailment (Clarke, 2006).

We can describe such models using a context theory $\langle A, A^*, \hat{\ }, \cdot \rangle$, where $\cdot$ is convolution in $L^1(A^*)$ and

$$\hat{x} = (1/2^{|x|}) \sum_{y \in \mathrm{Sub}(x)} e_y,$$

i.e. the context vector of a string is a weighted sum of its subsequences. Under this context theory $\hat{x} \leq \hat{y}$, i.e. x completely entails y if x is a subsequence of y.

Many variations on this context theory are possible, for example using more complex mappings to $L^1(A^*)$. The context theory can also be adapted to incorporate a measure of lexical overlap between strings, an approach that, although simple, performs comparably to more complex techniques in tasks such as recognising textual entailment (Dagan et al., 2005)

3.2 Lexical Entailment Model

Glickman and Dagan (2005) define their own model of entailment and apply it to the task of recognising textual entailment. They estimate entailment between words based on occurrences in documents: they estimate a *lexical entailment probability* $\mathrm{LEP}(x, y)$ between two terms x and y to be

$$\mathrm{LEP}(x, y) \simeq \frac{n_{x,y}}{n_y}$$

where n_y and $n_{x,y}$ denote the number of documents that the word y occurs in and the words x and y both occur in respectively.

We can describe this using a context theory $\langle A, D, \hat{\ }, \cdot \rangle$, where D is the set of documents, and

$$\hat{x}(d) = \begin{cases} 1 & \text{if } x \text{ occurs in document } d \\ 0 & \text{otherwise.} \end{cases} .$$

In this case the estimate of $\mathrm{LEP}(x, y)$ coincides with our own degree of entailment $\mathrm{Ent}(x, y)$.

There are many ways in which the multiplication $\cdot$ can be defined on $L^1(D)$. The simplest one defines $e_d \cdot e_f = e_d$ if $d = f$ and $e_d e_f = 0$ otherwise. The effect of multiplication of the context vectors of two strings is then set intersection:

$$(\hat{x} \cdot \hat{y})(d) = \begin{cases} 1 & \text{if } x \text{ and } y \text{ occur in document } d \\ 0 & \text{otherwise.} \end{cases}$$

Model	**Accuracy**	**CWS**
Dirichlet (10^6)	0.584	0.630
Dirichlet (10^7)	0.576	0.642
Bayer (MITRE)	0.586	0.617
Glickman (Bar Ilan)	0.586	0.572
Jijkoun (Amsterdam)	0.552	0.559
Newman (Dublin)	0.565	0.6

Table 1: Results obtained with our Latent Dirichlet projection model on the data from the first Recognising Textual Entailment Challenge for two document lengths $N = 10^6$ and $N = 10^7$ using a cutoff for the degree of entailment of 0.5 at which entailment was regarded as holding. CWS is the confidence weighted score — see (Dagan et al., 2005) for the definition.

Glickman and Dagan (2005) do not use this measure, possibly because the problem of data sparseness makes it useless for long strings. However the measure they use can be viewed as an approximation to this context theory.

We have also used this idea to determine entailment, using latent Dirichlet allocation to get around the problem of data sparseness. A model was built using a subset of around 380,000 documents from the Gigaword corpus, and the model was evaluated on the dataset from the first Recognising Textual Entailment Challenge; the results are shown in Table 1. In order to use the model, a document length had to be chosen; it was found that very long documents yielded better performance at this task.

3.3 Representing Taxonomies

In this section we describe how the relationships described by a taxonomy, the collection of **is-a** relationships described by ontologies such as WordNet (Fellbaum, 1989), can be embedded in the vector lattice structure that is crucial to the context-theoretic framework. This opens up the way to the possibility of new techniques that combine the vector-based representations of word meanings with the ontological ones, for example:

- **Semantic smoothing** could be applied to vector based representations of an ontology, for example using distributional similarity measures to move words that are distributionally similar closer to each other in the vector space. This type of technique may allow the benefits of vector based techniques and ontologies to be combined.

- **Automatic classification:** representing the taxonomy in a vector space may make it easier to look for relationships between the meanings in the taxonomy and meanings derived from vector based techniques such as latent semantic analysis, potentially aiding in classifying word meanings in a taxonomy.

- The new vector representation could lead to new measures of **semantic distance**, for example, the L^p norms can all be used to measure distance between the vector representations of meanings in a taxonomy. Moreover, the vector-based representation allows ambiguity to be represented by adding the weighted representations of individual senses.

We assume that the **is-a** relation is a partial ordering; this is true for many ontologies. We wish to incorporate the partial ordering of the taxonomy into the partial ordering of the vector lattice. We will make use of the following result relating to partial orders:

Definition 2 (Ideals)**.** *A* lower set *in a partially ordered set S is a set T such that for all $x, y \in S$, if $x \in T$ and $y \leq x$ then $y \in T$.*

The principal ideal generated by an element x *in a partially ordered set S is defined to be the lower set $\downarrow(x) = \{y \in S : y \leq x\}$.*

Proposition 3 (Ideal Completion)**.** *If S is a partially ordered set, then $\downarrow(\cdot)$ can be considered as a function from S to the powerset 2^S. Under the partial ordering defined by set inclusion, the set of lower sets form a complete lattice, and $\downarrow(\cdot)$ is a completion of S, the* ideal completion.

We are also concerned with the probability of concepts. This is an idea that has come about through the introduction of "distance measures" on taxonomies (Resnik, 1995). Since terms can be ascribed probabilities based on their frequencies of occurrence in corpora, the concepts they refer to can similarly be assigned probabilities. The probability of a concept is the probability of encountering an instance of that concept in the corpus, that is, the probability that a term selected at random from the corpus has a meaning that is subsumed by that particular concept. This ensures that more general concepts are given higher probabilities, for example if there is a most general concept (a top-most node in the taxonomy, which may correspond for example to "entity") its probability will be one, since every term can be considered an instance of that concept.

We give a general definition based on this idea which does not require probabilities to be assigned based on corpus counts:

Definition 4 (Real Valued Taxonomy)**.** *A real valued taxonomy is a finite set S of* concepts *with a partial ordering $\leq$ and a positive real function p over S. The* measure *of a concept is then defined in terms of p as*

$$\hat{p}(x) = \sum_{y \in \downarrow(x)} p(y).$$

The taxonomy is called probabilistic *if $\sum_{x \in S} p(s) = 1$. In this case $\hat{p}$ refers to the* probability of a concept.

Thus in a probabilistic taxonomy, the function p corresponds to the probability that a term is observed whose meaning corresponds (in that context) to that concept. The function $\hat{p}$ denotes the probability that a term is observed whose meaning in that context is subsumed by the concept.

Note that if S has a top element I then in the probabilistic case, clearly $\hat{p}(I) = 1$. In studies of distance measures on ontologies, the concepts in S often correspond to senses of terms, in this case the function p represents the (normalised) probability that a given term will occur with the sense indicated by the concept. The top-most concept often exists, and may be something with the meaning "entity"—intended to include the meaning of all concepts below it.

The most simple completion we consider is into the vector lattice $L^1(S)$, with basis elements $\{e_x : x \in S\}$.

Proposition 5 (Ideal Vector Completion)**.** *Let S be a probabilistic taxonomy with probability distribution function p that is non-zero everywhere on S. The function ψ from S to $L^1(S)$ defined by*

$$\psi(x) = \sum_{y \in \downarrow(x)} p(y) e_y$$

is a completion of the partial ordering of S under the vector lattice order of $L^1(S)$, satisfying $\|\psi(x)\|_1 = \hat{p}(x)$.

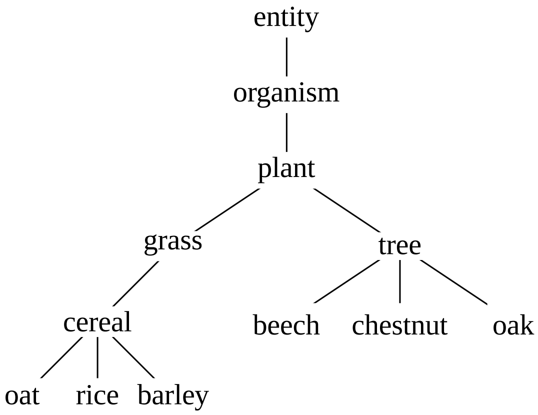


Figure 2: A small example taxonomy extracted from WordNet (Fellbaum, 1989).

Proof. The function ψ is clearly order-preserving: if $x \leq y$ in S then since $\downarrow(x) \subseteq \downarrow(y)$, necessarily $\psi(x) \leq \psi(y)$. Conversely, the only way that $\psi(x) \leq \psi(y)$ can be true is if $\downarrow(x) \subseteq \downarrow(y)$ since p is non-zero everywhere. If this is the case, then $x \leq y$ by the nature of the ideal completion. Thus ψ is an order-embedding, and since $L^1(S)$ is a complete lattice, it is also a completion. Finally, note that $\|\psi(x)\|_1 = \sum_{y \in \downarrow(x)} p(y) = \hat{p}(x)$. □

This completion allows us to represent concepts as elements within a vector lattice so that not only the partial ordering of the taxonomy is preserved, but the probability of concepts is also preserved as the size of the vector under the L^1 norm.

3.4 Representing Syntax

In this section we give a description link grammar (Sleator and Temperley, 1991) in terms of a context theory. Link grammar is a lexicalised syntactic formalism which describes properties of words in terms of *links* formed between them, and which is context-free in terms of its generative power; for the sake of brevity we omit the details, although a sample link grammar parse is show in Figure 3.

Our formulation of link grammar as a context theory makes use of a construction called a *free inverse semigroup*. Informally, the free inverse semigroup on a set S is formed from elements of S and their inverses, $S^{-1} = \{s^{-1} : s \in S\}$, satisfying no other condition than those of an inverse semigroup. Formally, the free inverse semigroup is defined in terms of a congruence relation on $(S \cup S^{-1})^*$ specifying the inverse property and commutativity of idempotents — see (Munn, 1974) for details. We denote the free inverse semigroup on S by $\mathrm{FIS}(S)$.

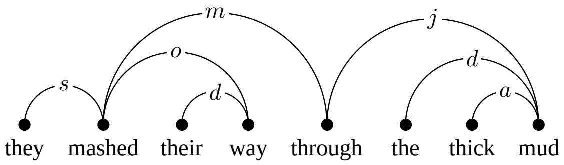


Figure 3: A link grammar parse. Link types: s: subject, o: object, m: modifying phrases, a: adjective, j: preposition, d: determiner.

Free inverse semigroups were shown by Munn (1974) to be equivalent to *birooted word trees*. A birooted word-tree on a set A is a directed acyclic graph whose edges are labelled by elements of A which does not contain any subgraphs of the form $\bullet \xrightarrow{a} \bullet \xleftarrow{a} \bullet$ or $\bullet \xleftarrow{a} \bullet \xrightarrow{a} \bullet$, together with two distinguished nodes, called the start node, □ and finish node, ○.

An element in the free semigroup $\mathrm{FIS}(S)$ is denoted as a sequence $x_1^{d_1} x_2^{d_2} \ldots x_n^{d_n}$ where $x_i \in S$ and $d_i \in \{1, -1\}$.

We construct the birooted word tree by starting with a single node as the start node, and for each i from 1 to n:

- Determine if there is an edge labelled x_i leaving the current node if $d_i = 1$, or arriving at the current node if $d_i = -1$.

- If so, follow this edge and make the resulting node the current node.

- If not, create a new node and join it with an edge labelled x_i in the appropriate direction, and make this node the current node.

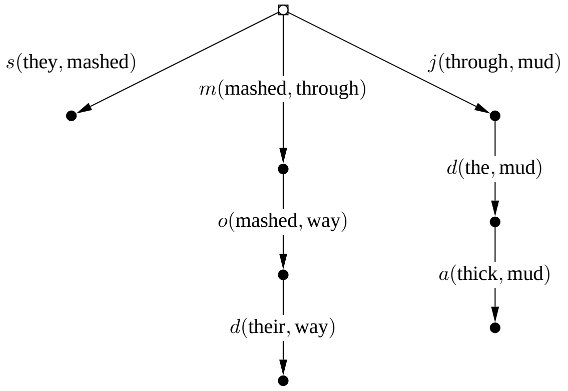
The finish node is the current node after the n iterations.

The product of two elements x and y in the free inverse semigroup can be computed by finding the birooted word-tree of x and that of y, joining the graphs by equating the start node of y with the finish node of x (and making it a normal node), and merging any other nodes and edges necessary to remove any subgraphs of the form $\bullet \xrightarrow{a} \bullet \xleftarrow{a} \bullet$ or $\bullet \xleftarrow{a} \bullet \xrightarrow{a} \bullet$. The inverse of an element has the same graph with start and finish nodes exchanged.

We can represent parses of sentences in link grammar by translating words to syntactic categories in the *free inverse semigroup*. The parse shown earlier for “they mashed their way through the thick mud” can be represented in the inverse semigroup on $S = \{s, m, o, d, j, a\}$ as

$$ss^{-1}modd^{-1}o^{-1}m^{-1}jdaa^{-1}d^{-1}j^{-1}$$

which has the following birooted word-tree (the words which the links derive from are shown in brackets):



Let A be the set of words in the natural language under consideration, S be the set of link types. Then we can form a context theory $\langle A, \mathrm{FIS}(S), \hat{}, \cdot \rangle$ where $\cdot$ is multiplication defined by convolution on $\mathrm{FIS}(S)$, and a word $a \in A$ is mapped to a probabilistic sum $\hat{a}$ of its link possible grammar representations (called *disjuncts*). Thus we have a context theory which maps a string x to elements of $L^1(\mathrm{FIS}(S))$; if there is a parse for this string then there will be some component of $\hat{x}$ which corresponds to an idempotent element of $\mathrm{FIS}(S)$. Moreover we can interpret the magnitude of the component as the probability of that particular parse, thus the context theory describes a probabilistic variation of link grammar.

3.5 Uncertainty in Logical Semantics

For the sake of brevity, we summarise our approach to representing uncertainty in logical semantics, which is described in full elsewhere. Our aim is to be able to reason with probabilistic information about uncertainty in logical semantics. For example, in order to represent a natural language sentence as a logical statement, it is necessary to parse it, which may well be with a statistical parser. We may have hundreds of possible parses and logical representations of a sentence, and associated probabilities. Alternatively, we may wish to describe our uncertainty about word-sense disambiguation in the representation. Incorporating such probabilistic information into the representation of meaning may lead to more robust systems which are able to cope when one component fails.

The basic principle we propose is to first represent unambiguous logical statements as a context theory. Our uncertainty about the meaning of a sentence can then be represented as a probability distribution over logical statements, whether the uncertainty arises from parsing, word-sense disambiguation or any other source. Incorporating this information is then straightforward: the representation of the sentence is the weighted sum of the representation of each possible meaning, where the weights are given by the probability distribution.

Computing the degree of entailment using this approach is computationally challenging, however we have shown that it is possible to estimate the degree of entailment by computing a lower bound on this value by calculating pairwise degrees of entailment for each possible logical statement.

4 Related Work

Mitchell and Lapata (2008) proposed a framework for composing meaning that is extremely general in nature: there is no requirement for linearity in the composition function, although in practice the authors do adopt this assumption. Indeed their “multiplicative models” require composition of two vectors to be a linear function of their tensor product; this is equivalent to our requirement of distributivity with respect to vector space addition.

Various ways of composing vector based representations of meaning were investigated by Widdows (2008), including the tensor product and direct sum. Both of these are compatible with the context theoretic framework since they are distributive with respect to the vector space addition.

Clark et al. (2008) proposed a method of composing meaning that generalises Montague semantics; further work is required to determine how their method of composition relates to the context-theoretic framework.

Erk and Pado (2008) describe a method of composition that allows the incorporation of selectional preferences; again further work is required to determine the relation between this work and the context-theoretic framework.

5 Conclusion

We have given an introduction to the context-theoretic framework, which provides mathematical guidelines on how vector-based representations of meaning should be composed, how entailment should be determined between these representations, and how probabilistic information should be incorporated.

We have shown how the framework can be applied to a wide range of problems in computational linguistics, including subsequence matching, vector based representations of taxonomies and statistical parsing. The ideas we have presented here are only a fraction of those described in full in (Clarke, 2007), and we believe that even that is only the tip of the iceberg with regards to what it is possible to achieve with the framework.

Acknowledgments

I am very grateful to my supervisor David Weir for all his help in the development of these ideas, and to Rudi Lutz and the anonymous reviewers for many useful comments and suggestions.

References

Y. A. Abramovich and Charalambos D. Aliprantis. 2002. *An Invitation to Operator Theory*. American Mathematical Society.

Charalambos D. Aliprantis and Owen Burkinshaw. 1985. *Positive Operators*. Academic Press.

Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. A compositional distributional model of meaning. In *Proceedings of the Second Symposium on Quantum Interaction, Oxford, UK*, pages 133–140.

Daoud Clarke. 2006. Meaning as context and subsequence analysis for textual entailment. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge*.

Daoud Clarke. 2007. *Context-theoretic Semantics for Natural Language: an Algebraic Framework*. Ph.D. thesis, Department of Informatics, University of Sussex.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Katrin Erk and Sebastian Pado. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*.

Christaine Fellbaum, editor. 1989. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts.

John R. Firth. 1957. Modes of meaning. In *Papers in Linguistics 1934–1951*. Oxford University Press, London.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, University of Michigan.

Oren Glickman and Ido Dagan. 2005. A probabilistic setting and lexical cooccurrence model for textual entailment. In *ACL-05 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*.

Zellig Harris. 1985. Distributional structure. In Jerrold J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

W. D. Munn. 1974. Free inverse semigroup. *Proceedings of the London Mathematical Society*, 29:385–404.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453.

Daniel D. Sleator and Davy Temperley. 1991. Parsing english with a link grammar. Technical Report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004, Geneva, Switzerland.*

Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Proceedings of the Second Symposium on Quantum Interaction, Oxford, UK*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Macmillan, New York. G. Anscombe, translator.