Computational Linguistic Models of Mental Spaces

Cliff O'Reilly

A thesis submitted for the degree of MSc By Dissertation

Department of Computer Science and Electronic Engineering

University of Essex

Date of submission: October 2014

Abstract

In this report we describe a computational linguistic model of mental spaces. We take theories from cognitive science as inspiration and, using the FrameNet database, construct a model upon which we execute a number of experiments.

Our underlying assumption is that, in order to develop computer systems that have near-human capacities for natural language processing, those systems will need to model cognitive processes. Gilles Fauconnier's theory of Mental Spaces provides a detailed background of partitioned semantic relations. These relationships can be constrained by Frames and Scripts. We use pre-existing computer tools to develop a model that mimics this framework. Fauconnier's and Turner's work on Conceptual Integration and current theories of dynamic systems are further inspiration for a model of conceptual integration using Latent Dirichlet Allocation, a topic modelling algorithm.

We choose three experiments with which to validate the usefulness of this approach. Our first experiment investigates text classification using the Full Text corpus within FrameNet. Our second experiment uses the corpora supplied for the SemEval Textual Semantic Similarity Task in order to validate the hypothesis that mental space networks are related to semantic similarity. The third experiment in this report investigates the Blending model and the hypothesis that this is related to the style of the document text.

The results for these experiments were mixed. We are pleased with some high Micro F1 scores (0.9), but disappointed that overall the results are not conclusive. We describe the analysis of the outcomes and also the drawbacks of our methods.

Finally we explain our thoughts on how these models could be improved and extended by learning lessons from our work and also including other work and approaches.

Contents

Chapter 1

Introduction

The approach taken in this research is one that uses theories from cognitive science and linguistics as inspiration to create a computer model that can mimic human-level language processing. This is an ambitious aim, but in this report it will be shown that there is potential in this methodology to improve computational linguistics systems.

The background to this report derives from the idea that artificial intelligence applications need world knowledge and context in order to provide mechanisms to calculate or capture the full meaning from language.

> "The key to building more powerful AI applications is to model the world knowledge and the linguistic and other basic abilities that people bring to bear. We now know that these abilities can not be fully expressed in abstract formalisms, but require models that map onto human biology and behaviour. Cognitive Science is the field that is best placed to unite the theory and applications of intelligence." [Feldman, 2007]

> "language does not represent 'meaning': language prompts for the construction of meaning in particular contexts with particular cultural models and cognitive resources it draws heavily on 'backstage' cognition" [Fauconnier and Turner, 2002]

> "Meaning Potential is the essentially unlimited number of ways in which an expression can prompt dynamic cognitive processes, which include conceptual connections,

mappings, blends, and simulations." [Fauconnier and Turner, 2002]

1.1 Pragmatics

"Why don't people just say what they mean?" [Thomas, 2014]

"It is possible that natural language has only syntax and pragmatics" [Chomsky, 1995]

Pragmatics is a field of linguistics concerning the meaning of natural language. Unlike Semantics, Pragmatics investigates how context affects the meaning of language in complex ways such as in social interaction. Pragmatics is also concerned with pre-existing knowledge of individuals and intention. In fact, in any of the almost infinite ways in which we communicate, we bring to the words more than the words themselves convey. Pragmatics also investigates the complex ways that words can have an impact on the world, for example with the theory of Speech Acts developed by J.L. Austin. Our research is interested in this area because our goal is to computationally record, describe and use the contextual information that humans use so effortlessly to give meaning to the world. Fauconnier's Mental Spaces theory investigates the structure of language, from a cognitive perspective that involves the pragmatic content of situations. Fauconnier's work on Mental Spaces develops the idea of Pragmatic Functions. Often called Connectors, these theoretical links connect entities in a mental space network, for example a situation in which a dog and a man interact would relate the dog and the man by a Pragmatic Function which links the two. Fauconnier contrasts the partitioning method necessary for the metaphysical Possible Worlds idea [Hintikka, 1962; Kripke, 1963] with a cognitive partitioning scheme [Fauconnier, 1994]. This drives a key concept for our research: that by partitioning the pragmatic information that is evoked by a text in a computational manner, we can develop beneficial language applications.

In order to derive a computer model, we use tools based on the theory of Frame Semantics, specifically FrameNet, which is a database of semantic relationships. These relationships extend into the contextual by way of the highly complex and interrelated database of frames and their relations. A FrameNet frame is a "script-like conceptual structure that describes a particular

type of situation, object, or event along with its participants and props" [Ruppenhofer et al., 2006]. We populate into the model, the evoked and related frames, participants and props, in order to capture many possible pragmatic interpretations of the sentence.

> "Language forms do not 'carry' information; they latch on to rich pre-existent networks in the subjects' brains and trigger massive sequential and parallel activations" [Fauconnier, 2004]

In order to compute meaningful outputs to these massive activations, we use machine learning algorithms over the evoked and related elements. Specifically, we use the topic modelling mixture model Latent Dirichlet Allocation.

1.2 Formalism and Dynamism

> "If concept formation can be explained by facts of nature, shouldn't we be interested, not in grammar, but rather in what is its basis in nature?" [Wittgenstein, 1953]

> "Cognition is a dynamic process, continually changing over time" [Prinz and Barsalou, 2014]

A modern trend in brain science is concerned with an embodied, situated or dynamical systems approach. This recent method "focuses on concrete action and emphasises the way in which an agent's behaviour arises from the dynamical interaction between its brain, its body and its environment." [Beer, 2014] Rather than a formalist or representational view of information processing, "a dynamical system is any system that evolves over time in a law-governed way" [Bermúdez, 2014]

Formal tools for understanding and manipulating language, such as parsing algorithms, are important, however we believe that this newer approach to information processing is potentially even more useful. From the perspective of natural language, we believe that the brain conforms to a mode more akin to a dynamic system than a formal or representational one.

The research described in this report uses the idea of dynamic systems to develop the output of the computer model we create. Our computer system is formal at first - using the output of

language parsing and populating a set of delineated mental space approximations that represent words, and evoked entities etc. This approach is representational. It contains rules and structures that may run counter to modelling a dynamic, neural system. However, we extend this by attempting to create a matrix of Meaning Potential with which to mimic neural processes. We then, via Latent Dirichlet Allocation and machine learning algorithms, restrict the resultant output of the computer system. In essence this is a simulation of a dynamic system.

1.3 Statistical models

We also assert in this work that statistical modelling is crucial to create computer systems that can model cognition effectively. The sheer volume of information, and therefore computer data, that surrounds even simple sentences must be filtered and manipulated in a goal-directed way such that the outputs are fit for our use.

> "A more radical argument for probability as part of scientific understanding of language is that human cognition is probabilistic and that language must therefore be probabilistic too since it is an integral part of cognition" [Manning and Schütze, 1999]

We use lexical cues and semantic database lookups to generate a multi-dimensional meaning space that correlates to the meaning potential from text. We further extend this by attempting to model emergent structures via the Latent Dirichlet Allocation algorithm.

1.4 The problem

As previously stated, we believe that the goal of human-level language processing is unachievable without addressing how humans process language at a cognitive level. While significant progress has been made in many areas of artificial intelligence research, there is still a long way to go before we can say that machines can process language in a similar capacity to humans.

Advances in cognitive science, and specifically cognitive linguistics, have progressed many theories that explain how language is processed. We have attemped to take a leading theory in this field - Mental Spaces theory - and develop a computer modelling method which investigates the benefits to a number of known problems. We also make use of FrameNet, which is a database

of semantic frames. By processing and partitioning textual data according to our interpretation of these theories we hope to mimic cognitive processing of language.

1.5 The structure of this thesis

The structure of this thesis is as follows:

- Chapter 1: Introduction - we introduce the ideas and motivations for the research
- Chapter 2: Background - we describe the cognitive science theories, the statistical tools, and the computer model developed in this project
- Chapter 3: Experiments - we discuss, in detail, the three experiments undertaken in our research. We also present the results for each experiment and an analysis of each one
- Chapter 4: Discussion - we take each experiment in turn and analyse the results in the context of what we were attempting to achieve
- Chapter 5: Conclusion & Future Work - we discuss our interpretation of the experimental results and present our thoughts on the drawbacks. We then present our ideas for further work on this project.

Chapter 2

Background

In the Introduction, we argue that to construct computers that can use natural language in a similar way to humans, we need to look to fields such as cognitive science for theories that can be modelled. In this way the mechanisms that underlie human mental processing of language can be brought into action by computing systems. In this chapter, we further expand on these ideas regarding some of the theoretical background from cognitive science and computer science and also some of the background to the techniques used in the implementation of the models constructed during this research. We re-iterate that this is an ambitious goal and we recognise that this report goes only a limited way towards it. We begin by looking to the theory of Mental Spaces and Conceptual Blending, both of which develop from the field of Cognitive Linguistics.

2.1 Mental Spaces

The Cognitive Science theory of Mental Spaces was developed by Gilles Fauconnier in the 1980s [Fauconnier, 1994], seeded by the wealth of research undertaken during the 1970s into the cognitive basis of language. A mental space is a conceptual packet assembled for purposes of thought and action. It is represented as a bounded set of elements with neural correlates, rather than a continuous domain. It is an abstract representation which, in the neural interpretation, is a set of activated neuronal assemblies with connections between spaces as coactivation-bindings (see [Yang et al., 2013], for neurological study of Blending). It "attempts to model the cogniser's

understanding of the world, not the world itself [and] contain[s] elements that include roles, values and relations recruited from various semantic frames" [Oakley and Hougaard, 2008]. Mental Spaces can be distinguished into types and can be constructed in many ways, for example when new information is given or discovered, and under the influence of cues in language. Types include Domain spaces, where world knowledge can reside; Space spaces, related to locations; and Time spaces, related to chronological events. Space-building linguistic cues might be changes in tense, conditionals or locative and temporal phrases, e.g. "In 1929" or "In London".

As an example, see Figure 2.2 where a simple set of mental spaces has been shown involving a hypothetical conversation between a modern philosopher and Immanuel Kant (The Debate with Kant network). The example sentence used is: "I'm claiming that reason is self-developing. Kant says that it's innate. I mention Neuronal Group Selection and he gives no answer." [Fauconnier and Turner, 2002]

Input spaces 1 and 2 contain elements related to Kant and to the modern philosopher and after the blending process a new mental space is constructed that contains references to the inputs, but also emergent structure.

Fauconnier develops the Mental Space theory by relating the components and elements to Frames. "They are ... structured by frames" and "we say that the mental space is framed and we call that organisation a 'frame'". Further, Fauconnier describes a situation with various participants. This relates exactly to a Semantic Frame (Frame Semantics) and therefore to the concept of a Frame in FrameNet: Commercial Event [Fillmore, 2006] and the Commerce_buy frame in FrameNet: "a mental space in which Julie purchases a coffee at Peet's coffee shop has individual elements that are *framed* by commercial transaction as well as by the subframe - highly important for Julie - of *buying a coffee at Peet's*" [Fauconnier and Turner, 2002].

Connections across and within spaces relate elements by Pragmatic Functions. These are relationships that are often complex and multi-dimensional and, by addressing semantic frame structures, can generate a very large network of inter-connections. In effect, by referencing and connecting between all the many mental spaces and frame-related contextual information, an enormous cognitive possible worlds network is generated. This is a dynamic and very large structure: "mental spaces are partial models of present, past, future, possible, impossible, or otherwise imagined states of affairs understood by the cognizer. They are not models of the

world; they are dynamic models of the moment-to-moment understanding of states of affairs" [Oakley, 2009]

2.2 Conceptual Integration (Blending)

A Conceptual Integration Network is a network of mental spaces that are inter-connected and interact in complex ways via connectors and rules. Connectors can be of various kinds such as psychological, cultural or pragmatic. Connectors and Counterparts link objects across spaces, for example:

- "He thinks" - a Mental Image Connector that links from reality to beliefs
- "In the picture" - Image Connector that links from models to pictures
- "In that movie" - Drama Connector that links from actors to characters

A blended mental space is an integration of received input projections from other mental spaces in the network and it develops emergent structure not available from the inputs alone. It operates under a set of constitutive and governing principles. In a Conceptual Integration Network (the blending situation) the various categories of mental space include (see Figure 2.1 for the basic diagram):

- **Input space** - the spaces that exist prior to the blend and any relationships and connectors
- **Generic space** - a single space that contains generic versions of the elements in the input spaces, e.g. if the input spaces both have human people represented then the generic space would have a 'human' object that is connected to all the input space objects that are human
- **Blend space** - these are the outputs of the blending process and contain new objects and connectors to existing objects in the input and generic spaces.

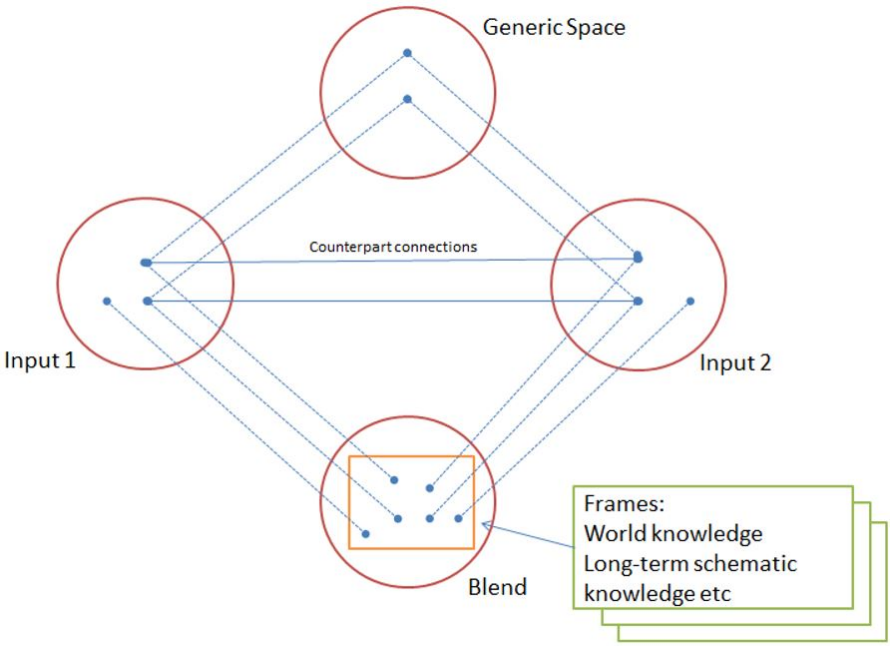

Figure 2.1: Conceptual Blending - the basic diagram (adapted from [Fauconnier and Turner, 2002])

Blending develops emergent structure not contained in the input spaces by a standard process:

- Composition – "new" objects are created in the blended spaces
- Completion – objects interact imaginatively/logically to "complete" the scenario
- Elaboration – simulated scenarios are "played out" to infer new objects and relationships

Related work on computational models of Blending have tended to focus either on a generative ([Goguen and Harrell, 2004], [Harrell, 2005]) or representational and algorithmic approach ([Veale and O'Donoghue, 2000]). This research's implementation of Conceptual Integration theory does not correspond exactly to the representational view of mental spaces as described previously. It is also not generative nor algorithmic. What we attempt is to mimic the general function of Blending via the process of using grouped, differentiated textual elements (and their associated correspondences, e.g. frame-related elements) as inputs to statistical mixture models. We theorise that this is an approximation to Blending, where the grouped texts approximate to mental spaces and the statistical models approximate to Blending. Our implementation is much

simplified and does not relate groups (what we call m-Frames) in a structured way, as in the Mental Spaces theory, however we discuss this as a possibility for future work.

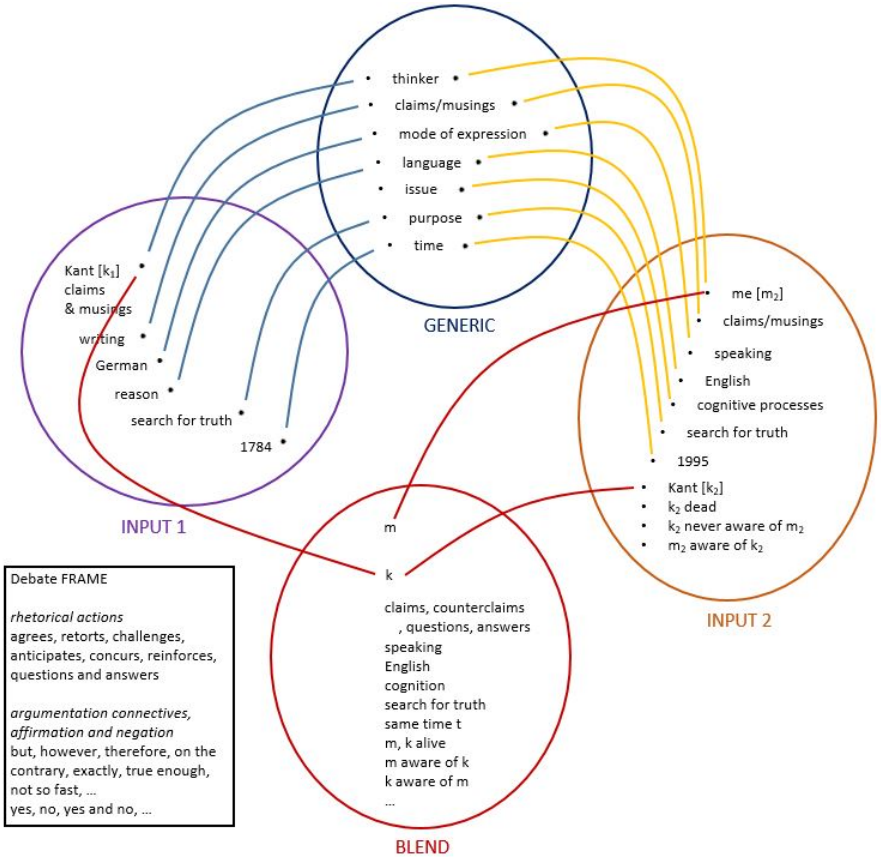

Figure 2.2: Example set of Mental Spaces and Blending (adapted from [Fauconnier and Turner, 2002])

2.3 Frame Semantics & Scripts

From the above, it can be seen that a model involving mental spaces requires semantic frames. The two work together in tandem, one providing relationships and referential framework (mental spaces theory) and the other filling in the concepts (frames). FrameNet facilitates a very good method to develop frame semantic elements from text. By using lexical matching (after usual pre-processing and parsing) it is possible to link words to frames and then by extension to fill out mental spaces with conceptual information related to the sentence, see Figure 2.3 for a

diagrammatic representation.

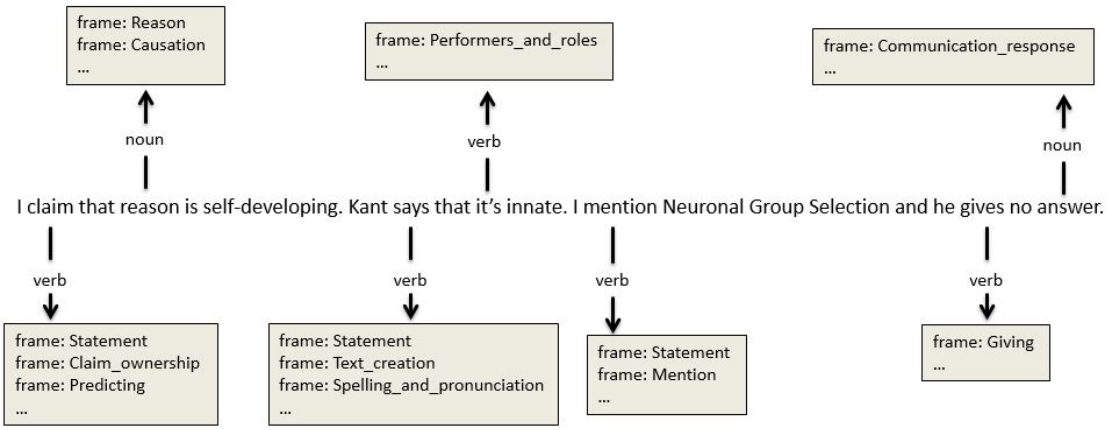

Figure 2.3: High level view of some of the frames evoked from the example sentence

Frame Semantics, developed mainly by Fillmore and Minsky in the 1970s and 80s, is a theory that says that "words represent categorisations of experience, and each of these categories is underlain by a motivating situation occurring against a background of knowledge and experience" [Fillmore, 2006]. In this theory a Frame is a system of categories linked to words. The motivating context is a collection of influences that humans have reason to be concerned with such as social manners, community history and practices, recent contextual cues. Meanings have internal structure which is determined relative to a background frame or scene; "to understand any one of them [a concept] you have to understand the whole structure [of concepts] in which it fits" [Fillmore, 2006].

Frames are often situational, for example the Commercial Transaction Frame, which consists of buyer, seller, goods and money elements etc (see Figure 2.4 [Hamm, 2007]). The connections between these words and specific or general situations are obvious (not all frames are like this).

VERB	BUYER	GOODS	SELLER	MONEY	PLACE
buy	subject	object	from	for	at
sell	to				
cost	indirect object	subject		object	at
spend	subject	on		object	at

Figure 2.4: Frame Semantics - Commercial Transaction Frame ([Hamm, 2007])

The ability of Frame Semantics to formalise contextual information in relation to text, especially when used with Mental Space theory, is crucial when attempting to process the semantics of natural language. When analysing words we need to see the bigger picture if we are to get to the meaning(s) that can be attributed to those words: "Frame Semantics, as a common, largely language-independent word sense and role inventory, holds great promise for the cross-lingual analysis and application of lexical semantic information." [Burchardt et al., 2006]

We recognise, too, the work of Schank and Abelson on Scripts. Similar to Frame Semantics, the theory of scripts relates scenes and plans to particular contexts, where the chronology or sequence is important - "A script, as we use it, is a structure that describes an appropriate sequence of events in a particular context" [Schank and Abelson, 1975]. Whereas, then, Frame Semantics defines a "situation against a background of knowledge or experience", a Script could be a "pre-determined, stereotyped sequence of actions that define a well-known situation" [Schank and Abelson, 1975]. From the perspective of our project there is overlap between these theories, however they treat the context in subtly different ways. Our research doesn't distinguish between them except that we utilise FrameNet which is more connected to Frame Semantics. We predict that, to be as effective as possible, future iterations of this model would benefit from scripts being integrated somehow so that mental spaces can be organised, for example chronologically, in relation to pre-defined sequences.

2.4 FrameNet

Our intention is to model mental spaces. As we have discussed in previous sections, we require a method to model semantic frames in order to supply contextual organisation. We want to create approximations to mental spaces and in order to structure the resultant network of data we need a data source for the context and structure. We are fortunate to be able to use FrameNet as a source of data.

FrameNet[1] [Ruppenhofer et al., 2006] is a lexical database of English that has been used extensively in computational Natural Language research. Derived from the concept of a Frame (from Frame Semantics), it is composed of a dictionary of word senses that centre on Frames. Frames are made up of Frame Elements, which can be Core or Non-Core and also a set of Frame to Frame relations. Words that can evoke a frame are called Lexical Units. For example the words fry, bake and boil can evoke the frame Apply_heat that contains frame elements such as Cook and Heating_Instrument etc.

A frame can be considered as a "script-like conceptual structure that describes a particular type of situation, object or event along with its participants and props" [Ruppenhofer et al., 2006]. Frame elements describe the semantic roles of the frame, whereas Lexical Units are the words that evoke the frame. There are many different Frame Elements including Location, Theme, Degree, Duration etc.

Frame to frame relations capture the relationships between frames in a structured way. This is described in Tables 2.1 [Ruppenhofer et al., 2006] and 2.2. This framework allows a semantic network to be described automatically by referencing first the evoked frames and then to consider related frames. This is a corollary for the idea that, as we process language, we do so via a network of semantic relationships. We can mimic this semantic network to a degree, by generating a FrameNet network. For example, the frame Arraignment has a frame-frame relation with the Criminal_Process frame as a **Subframe of** relationship. The **Precedes** relation includes the Arrest frame and the **Preceded by** includes the Trial frame.

[1] `https://framenet.icsi.berkeley.edu/fndrupal/home`

Relation	Sub	Super
Inheritance	Child	Parent
Perspective_on	Perspectivized	Neutral
Subframe	Component	Complex
Precedes	Later	Earlier
Inchoative_of	Inchoative	State
Causative_of	Causative	Inchoative/State
Using	Child	Parent

Table 2.1: FrameNet frame to frame relation ([Ruppenhofer et al., 2006])

Frame-frame relation	Example
Is inherited by	**Mention** "Is inherited by" **Indicating**
Perspectivized on	**Drop_in_on** "Perspective on" **Visit_host_arrival**
Uses	**Abusing** "Uses" **Cause_harm**
Used by	**Diversity** "Used by" **Delimitation_of_diversity**
Has subframe	**Activity** "Has subframe" **Activity_start**
Causative	**Emitting** "Caustive of" **Emanating**
Preceded by	**Trial** "Is preceded by" **Arrest**
Inherits from	**Absorb_Heat** "Inherits from" **Becoming**
Precedes	**Arrest** "Precedes" **Arraignment**
Subframe of	**Arraignment** "Subframe of" **Criminal_process**

Table 2.2: FrameNet frame to frame relation examples

2.5 Latent Dirichlet Allocation

As we have discussed so far in this chapter, the cognitive sciences provide theories that describe the way humans process language, at a cognitive level. In this project we attempt to create a model based on some of these theories and also extend this formal approach into a dynamic one. The purpose of this is to blend together the various input data in a statistical analysis that mimics the conceptual integration that occurs in the brain. Subsequent to creating a model of mental spaces (our approximations) we want to develop this by calculating a series of statistical models. Our intuition is that our mental space approximations are partitioned in such a way as to develop relationships in a contextual domain both in tandem with, and as a comparison to, statistical models exercised over words in sentences. We chose Latent Dirichlet Allocation for this task since it provided an unsupervised learning algorithm (we don't have training data with which to calculate under supervised algorithms) and is becoming widely used in computational

linguistics. It also provided a mixture model that matches our dynamic aims and has been shown to perform very well at topic discovery.

Latent Dirichlet Allocation, LDA, is a Generative Probabilistic Model under the rubric of Topic Models – a suite of algorithms that aim to discover thematic information [Blei et al., 2003]. The purpose of this statistical model is to analyse discrete datasets such as text corpora, but can be used with other domains, e.g. images and genetic data [Blei et al., 2010]. In comparison with earlier techniques like tf-idf, Latent Semantic Indexing (LSI), and Probabilistic Latent Semantic Indexing (pLSI), LDA utilises the exchangeability principle of words and documents and, as per de Finetti, considers mixture models to capture intra-document statistical structure. In the LDA model, the exchangeability principle can be seen as meaning that elements are independent and identically distributed and conditioned by underlying latent parameters. Further, the elements can be words in sentences, but also extended to, for example, n-grams or paragraphs.

The intuition behind LDA is that documents exhibit multiple topics. The generative process assumes that there exists a posterior distribution over the hidden random variables (the topic structure). That is then calculated from a joint probability distribution of those hidden variables and the observed variables (the vocabulary of words). More formally:

$$
\begin{aligned}
&p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\
&= \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \\
&\left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)
\end{aligned}
$$

Figure 2.5: LDA formula

where K=number of topics; D=number of documents; $\beta_{1:K}$ – topics themselves; $\Theta_{1:D}$ – topic proportions (has dimension K); $z_{1:D}$ – topic assignments; $w_{1:D}$ – observed words. [Blei et al., 2010]

Another way of analysing LDA is with a graphical model, e.g. in Figure 2.6.

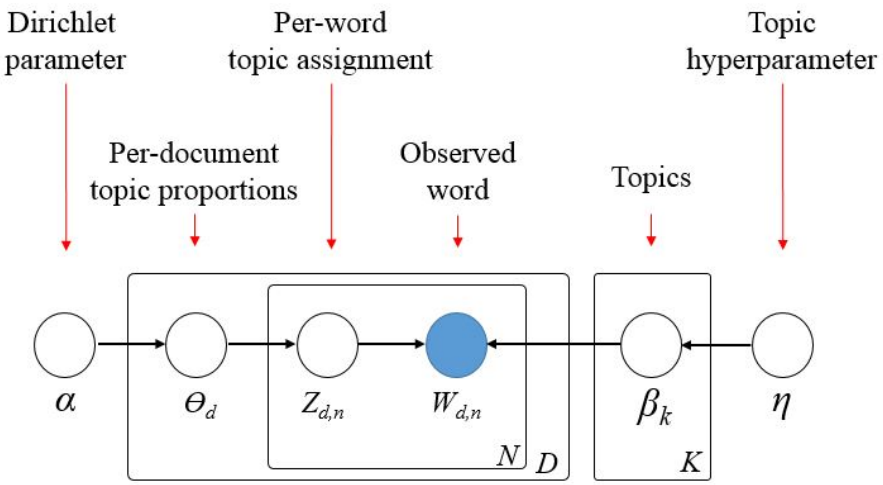

Figure 2.6: LDA graph

In this diagram the same algebraic elements are presented where the boxes are "plates" that represent replication of elements. The inner plate is the repeated choice of topics and words within a document and the outer plate represents documents. The Dirichlet parameter (α) controls the mean shape and sparseness of the topic distribution.

Each topic is considered to be a distribution over a fixed vocabulary. The algorithm that calculates the topic distribution has no background information about the topics: we infer the hidden topic structure by calculating the joint distribution of every possible instantiation of hidden topic structure, however for large data sets this is too large a calculation. Instead we approximate the posterior distribution. There are two approaches to this:

- Sampling-based – we collect samples from the posterior in order to approximate, e.g. Gibbs sampling using Markov-chains
- Variational-based – we posit a distribution over the hidden structure and find the member that is closest to the posterior (in this case it becomes an optimisation problem).

Each method has benefits and the use depends on the problem to be addressed. The process can be shown as a series of steps:

i For each document d, draw a topic mixture Θd from Dir(Θd; αn)

ii For each topic t, draw a distribution over words ϕt from Dir(ϕt; βm)

iii For each position i in document d:

iv Draw a topic z_i from Θd

v Draw a word w_i from ϕz_i

All the documents in the corpus share the same topics and exhibit each topic in different proportions related to the probability of each element in the topic. An assumption that we work under in this research is that, since words are related to all topics via probability distributions, topic-to-topic distributions can be generated from the set of word-to-topic distributions.

The usual method of evaluation is to hold out a test segment of data and then run the model against this held out section in order to validate the effectiveness. By running various models we can determine the best one by using these results. One of the problems with LDA at the moment is that, in certain domains, where supervised learning methods are not possible, there is difficulty in evaluating the models effectively. The best method for selecting the most appropriate model for a task is currently an open problem. The advantages of LDA over other models are that it can be readily "embedded in a more complex model" (not possible with LSI) [Blei et al., 2003]; also the probabilistic nature of the topic discovery is very useful:

> "Representing the content of words and documents with probabilistic topics has one distinct advantage over a purely spatial representation. Each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms." [Steyvers and Griffiths, 2007]

But most important for this project is that the model can be extended and augmented with various techniques. There are a number of assumptions that LDA uses and by relaxing these we can obtain different results. The bag of words assumption is that the word order is not important to topic classification (see [Wallach, 2006] and [Griffiths et al., 2004] for examples of relaxation and extension of this assumption by extending the model to include a bigram language model and Hidden Markov Models, respectively). Similarly, the order of documents is assumed not to matter for standard LDA models, however by ordering the documents, a richer posterior topical structure can be obtained that is, for example, dynamic over time [Blei and Lafferty,

2006]. There is also an assumption that the number of topics is fixed and known. By using Markov-chain Monte Carlo sampling schemes for posterior inference with hierarchical Dirichlet processes, [Teh et al., 2006] are able to determine the number of topics before the main algorithm starts and also allow new documents to suggest new topics.

We make use of a variant of LDA, called Labeled LDA (L-LDA) ([Ramage et al., 2009]), which is different from the standard LDA algorithm in that the topics chosen are constrained to a set of pre-defined topics. This enables the model to proceed with some supervision. It has been used as a generative model for labeled corpora, often with multiple labels per word. We use this variant to constrain the resulting topics to the supplied list of frames that we calculate as being evoked by sentences. See Figure 2.7 for a plate graph of the L-LDA algorithm.

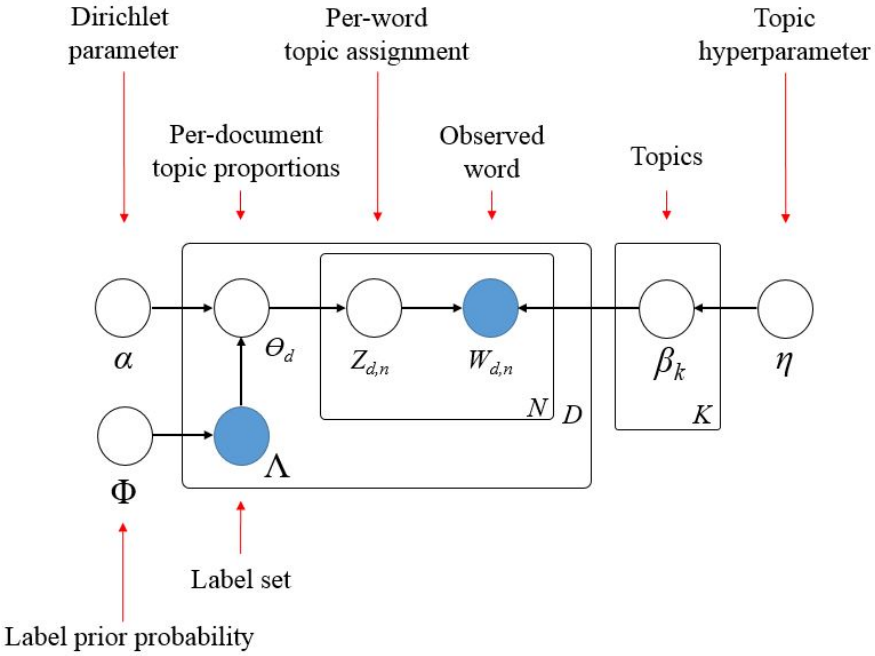

Figure 2.7: Labaled LDA graph

2.6 The Model

The aim of the model is to approximate a Mental Spaces Network (MSN) that is evoked by words and phrases in a text (although this could extend to any form of language or communication such as spoken, non-verbal etc). The intuition here is that it is not possible, computationally, to fully

represent meaning without using a model encompassing pragmatic complexity; without adding contextual and situational information to computer models of language it will not be possible for full processing of the meaning that humans develop when they process communications.

This model attempts to construct an approximate MSN that goes some way towards representing the actual, likely invoked elements. The model is limited to the dictionary of relationships (FrameNet) and the various corpora used. We are therefore careful not to assume too much into each MSN that is evoked. Each time we analyse the relationships of words and phrases we only can populate the model with a specific instance of a network. A real, human MSN may not be the same; it may be larger or smaller than the one(s) we assume. Fauconnier and Turner theorise that "mental spaces are small conceptual packets constructed as we think and talk and ... correspond to activated neuronal assemblies and linking between elements corresponds to some kind of neurobiological binding, such as co-activation.", and further, that "Meaning Potential is the essentially unlimited number of ways in which an expression can prompt dynamic cognitive processes, which include conceptual connections, mappings, blends, and simulations." [Fauconnier and Turner, 2002]. The real Mental Space network developed in the brain is likely to be a vast, almost incomprehensibly complicated interconnection of activated excitatory and inhibitory neurons and bundles of neurons, encompassing both long-term memory and working memory. This network would seem impossible to model. The important point for this project is that the computer system mimics the natural process and is only an approximation which we can incorporate into experiment, hopefully in order to improve computational linguistic applications.

The key area that we utilise from Mental Spaces theory is the partitioning and population of discrete collections of entities related to and evoked by words in a sentence. We create bounded assemblies that contain particular words or related elements from text. In effect this is simply a text file that contains comma-separated groups of words in columns that are used by the controlling Java program. The rules used to populate the mental space vary depending on the experiment being undertaken, but could include the nouns and the evoked frames - literally just the noun words and names of the evoked frames. We continue to populate these mental space approximations with elements from the FrameNet database, for example, further frames that are related to the initially-evoked frames along pre-defined relationship types, such as Inherited By.

We are using only the most rudimentary facet from the theory of Mental Spaces. We do

not extend our model to include the many ways in which mental spaces vary and interrelate. We explain in the Future Works section in this report that these further areas of the theory are possible to develop, but this was not included in the research so far.

The basic model, described above, is useful as a resource for analysis, for example in our Semantic Similarity experiment, but we also overlay statistical analysis on the model. By using LDA we "blend" together the elements from the MSN. This gives us an approximation to Conceptual Integration (Blending). Again we are conscious that this process is limited by the LDA algorithm and the necessary assumptions made by the nature of the information available to the model. However, the potential to provide further enhancements to computer applications of natural language is exciting.

Our approach only goes as far as representing a semantic network of the basic elements from the Mental Spaces theory. It does not include relationships between spaces nor differentiation between different kinds of mental spaces, for example, but it would be straightforward to do so given that these experiments are focussed on comparison. This is discussed further in the Future Work chapter. This initial simplicity is a significant drawback for a more generic model, but there are clear areas for extension that are also discussed in the Future Work section of this report. The purpose of this rudimentary model is to make an approximation of the contextual information around a sentence and do something useful with it.

Chapter 3

Experiments

In the previous chapter, we described the background and motivation for a computer model to be constructed that will encompass pragmatic information in order to provide a set of related information to be called upon in various experiments. In the model the initial pragmatic information comes from the FrameNet database, by way of the script-like conceptual structure that we query. Following the creation of the MSN, we extend the pragmatic information by "blending" the results via LDA topic discovery and analysis. In this chapter we present a technical description of the modelling approach, the experiments performed and the results obtained. We describe the basic model that underlies the various experiments undertaken, consisting of a multi-layered pre-processing platform upon which a series of experimental models are built. We give examples of the construction of the computer model and describe the relations to the theoretical background. We also discuss the three experiments that were undertaken. The first is a classification task using the FrameNet corpus, the second is a textual semantic similarity experiment using the corpus from the SemEval task, and the third experiment is an analysis of blending and writing style. After describing the basic approach we discuss each experiment in turn, along with the results obtained.

3.1 Extracting Mental Spaces from text - the Basic Approach

For a given text, structures from input sentences are analysed, in order to construct bounded collections of related entities. In effect this takes the form of organised bags-of-words collections that can be used in various tasks. The types of entities can include words, phrases, Names or FrameNet entities. As discussed in section 2.4, FrameNet categorises various elements according to their function. For example the frame, named **Time_vector**, is related to the Lexical Unit **before.prep** and the Frame Element of **Direction** etc. The bounded collections we call **m-Frames** and are grouped into different types. The basic m-Frames include collections of nouns, lexical units or named entities from the sentence. Using the FrameNet database, more complex m-Frames are put together that may contain the frames evoked by the sentence and, further, collections of related frames to that original evoked frame. These collections of related frames are separated into m-Frames for each Frame-Frame relation, e.g. Inchoative, Causative etc. Further still, the frames related to the original evoking frame can have their associated lexical units grouped into an m-Frame. For example the word, part-of-speech pair [bake, verb] evokes the frame Apply_Heat. The Apply_Heat frame is related to the Cooking_Creation frame by the Is Used By relationship. The Cooking_Creation frame is evoked by many lexical units, including [prepare, verb] and [concoct, verb] etc. The words Bake and Prepare and Concoct are all synonyms and this relationship could be discovered via other lexical databases. In this model, however, there is a semantic connection between the words that is related via a well-defined and queryable semantic network (FrameNet).

The Mental Space Networks that are constructed can become vast in size and dimension very quickly, for example a single sentence could evoke ten frames, each of which could be related to ten frames. Each of these one hundred frames could be referenced by ten Lexical Units and ten Frame Elements, giving a total of two thousand elements from a single sentence. This is only two levels of transition along the network, but there's no reason why three or more cannot be calculated. This will result in a large network that has many valid, but unlikely connotations for the sentence. The model is initially large in scope and in order to discover useful meaning it needs to be constrained or filtered.

Tables 3.1 and 3.6 show examples of m-Frame matrices for the word Claiming and also as applied to the Semantic Similarity problem.

3.2 Pre-processing

In this section we present the pre-processing mechanism. All the following experiments use the same pre-processing mechanism. The process is divided into phases during which a set of comma-separated variable (csv) files is produced. The csv files form the output and input of adjoining phases, Figure 3.1 shows a representation of the process.

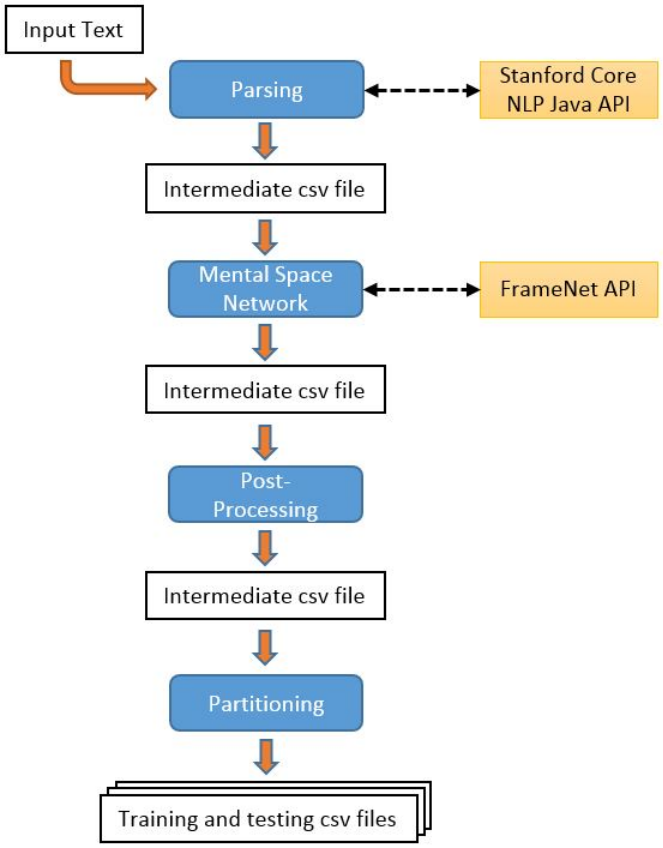

Figure 3.1: Pre-Processing data flow

This model and subsequent analysis is undertaken using the Java programming language, implemented in a suite of interrelated classes. The processing of input texts takes place via

the Java API implementation of the Stanford CoreNLP tool[1]. The input text file is tokenized, sentence-split, part-of-speech tagged and lemmatized. Named Entity Recognition, grammatical dependencies and Coreference occurrences are also analysed and the whole parse output saved as a csv file. This parsed file is then input into the next phase which uses the FrameNet API by Nils Reiter[2]. This API is a Java implementation of an XML reader, specific to extracting data from FrameNet. The FrameNet API does not provide disambiguation or computation in our model. In our case we use it purely to extract relations from FrameNet based on the data we supply. It facilitates the easy manipulation of the FrameNet database which is, in essence, a set of text files. We supply a lemma and part of speech pair to the API, which then returns the FrameNet Lexical Unit(s) that are related in the FrameNet model. As each word from the text, and its part of speech, is cross-referenced against the FrameNet database of lexical units (via the API), matches indicate a link to a FrameNet frame. For example the word **claiming** when lemmatized and tagged to **[claim - Verb]**, links to the FrameNet lexical unit **claim.v**, which, in turn, evokes the **Statement** frame. These evoked frames go to populate the FrameNet Frames m-Frame for the sentence.

Word	**FrameNet frame**	**Frame Elements**	**Frame relations**
claiming	Statement	Medium Place Epistemic_stance Depictive Iteration Message Manner Internal_cause Group Event_description Means Time Particular_iteration Degree Topic Frequency Addressee Occasion Containing_event Speaker	Recording Complaining Reveal_secret Telling Chatting Unattributed_information Attributed_information Adducing Judgment_communication Renunciation Communication
claiming	Claim_ownership	Claimant Beneficiary Role Property	Communication
claiming	Predicting	Place Eventuality Descriptor Time Accuracy Manner Evidence Speaker Time_of_Eventuality	Expectation

Table 3.1: Example m-Frame (partial) output for the word claiming

Following the discovery of frames, the pre-processing continues by adding Frame Elements

[1] `http://nlp.stanford.edu/software/corenlp.shtml`
[2] `http://www.cl.uni-heidelberg.de/trac/FrameNetAPI`

and Lexical Units for the originally-evoked frames and other frames that relate to the first frames, grouped by relationship type (Inchoative, Causative etc). An example output is shown in Table 3.1 and a high level view of the example sentence, as already seen, in Figure 2.3.

After adding m-frames, various further comma-separated variable files (csv) are output depending on the experiment. Each of these will be looked at in turn, below.

3.3 Experiment 1: Classification Task

3.3.1 Method

This experiment measures the ability of the Stanford topic modelling toolbox to classify unseen sentences. For this classification task we use the Full Text corpus that is associated with the FrameNet database. This corpus comprises 79 texts across various subjects, manually-annotated with evoked frames. There are 4026 sentences in the corpus. Within the annotation, each frame is related to the evoking word by sentence position. This enables a more fine-grained analysis, however, for this experiment, we have only used the relationship between evoked frame and the sentence. The fact that the corpus is manually-annotated gives us a good indicator to use for classifying text.

In this experiment we take the output of the pre-processing phase and run a Labeled Latent Dirichlet Analysis (L-LDA) model over it. This is an approximation to Blending (as mentioned in previous sections) that gives a probability distribution over groups of words (Topics) by providing a guiding set of Labels for each sentence. In effect it enables us to relate the Labels (the associated frames) with the sentence, probabilistically, and across the entire corpus.

Our hypothesis is that there is a relationship between the L-LDA distribution such that we can classify unseen texts with appropriate topics (frames).

As described previously, L-LDA allows us to restrict the probabilistic topic discovery to a set of associated labels. In this experiment these labels are initially the names of the annotated frames. We extend this, however, by varying the set of input labels across the m-Frames that were collated in pre-processing. The assumption is that we don't know what the best classifying features will be - it may be that frames alone are enough, however, perhaps a combination of

frames, nouns and subframes is better.

When undertaking this experiment we noticed that the annotated frames would sometimes be incomplete, i.e. that there are evoking words in the sentences whose frames are not in the manual annotation set. We also discovered that, due to the nature of the pre-processing work being done, that some manually-annotated frames were not included in our automated frame discovery process. In both cases we added the two sets together to form a complete set of evoked frames - some from the manual annotation and some from the automatic lexical matching, pre-processing phase.

After running the Java pre-processing modules on the input text, the Stanford Topic Modelling Toolbox[3] was used to create the L-LDA model.

In this scenario, repetition of words can have an impact on the LDA model so, for example, we don't remove duplicate m-Frame elements if they appear multiple times in the output document. The output word frequency is a consequence of the relationships discovered by the pre-processing and may be important. The fact that the same frame may be evoked multiple times in the same sentence could have relevance. The approach has been to leave words rather than filter them as they add to the mixture model of LDA.

At the end of the pre-processing phase the main output is split, in order to provide 10-fold cross validation.

The Stanford Topic Modelling Toolbox uses a Scala script to manipulate the input file and alter parameters of the model. We use most of the standard parameters and input processing variables, but vary the Term Smoothing between 0.01 and 0.5. We learn and then infer an L-LDA model on the csv files processed in the previous step. The LDA algorithm works in three stages:

1. Use the training dataset to learn a probability distribution, guided by the label set, and save the model

2. Use the training dataset to generate a per-sentence distribution over all topics and then to generate a per-label distribution over topics

3. Use the test dataset to generate a per-sentence distribution over all topics

[3] `http://nlp.stanford.edu/software/tmt/tmt-0.4/`

The first step of the L-LDA process outputs a distribution over documents (our input training sentences), quantifying how related each one is to each topic. Step 2 in this process is necessary in order to create a Gold Standard with which to compare the test data inference distribution. The training dataset is used to infer using the model generated in step 1. The per-label topic distributions gives us the relationship between labels and topics, from which we create a ranked list.

Since the distribution relates all topics to all documents (sentences) with varying probability, we need to arrange a cut-off probability level to reduce the number of results. We also realise that, similarly, the distribution across topics and labels includes all topics and labels and therefore a cut-off number of labels to associate with each topic is necessary. Varying these numbers causes the recall and precision to fluctuate respectively and so we chose numbers that seemed to be intuitive initially and then varied them to achieve an appropriate output.

Next we analyse the distribution data in order to compare the Gold Standard of classification with the Predicted Labels. In effect we have a multi-class, multi-label classification problem and therefore to analyse the effectiveness we use Micro-F1 and Macro-F1 scoring. We have a classification prediction score for each iteration of the experiment - for each m-Frame used as the label set, for each Term Smoothing parameter fed to the L-LDA algorithm, for each label cut-off level, and for each cut-off probability. This is averaged over the 10-fold cross validation set and gives us the results detailed in the next section of this report.

With multi-label classification, there are two methods used in this report to calculate the average across the sets of data: micro-average and macro-average (also known as Micro and Macro F1). In Micro F1, precision and recall are obtained by summing over all individual decisions, whereas in Macro F1, precision and recall are calculated "locally" for each category and then "globally" by averaging over the results of the different catagories. We included both methods of calculation because "These two methods may give quite different results" [Sebastiani, 2002].

3.3.2 Experiment Summary

In summary, we use the Stanford Topic Modelling Toolbox to train and then infer the topics on a set of inputs. The inputs are the sentences from the FrameNet full text corpus and the topic-constraining labels, which are derived from the pre-processing algorithms. This experiment was executed many times with varying input labels, such as the associated evoked frames. We extended the experiment by using the various sets of related frames as labels in input files, and even used the part of speech tags and part of speech group. An example input record is shown in Table 3.2. In Table 3.3, we show how the m-Frame relates to the input text. As described previously, what we call the m-Frame is a bag of words construction, intended to approximate a mental space. Our experiments create many possible versions of an m-Frame in order to calculate the many possible interpretations of a sentence.

ID	Corpus File	Sentence	Labels
1	ANC__110CYL067.xml	Your contribution to Goodwill will mean more than you may know.	Giving Goal Purpose Increment Awareness Likelihood
2	ANC__110CYL067.xml	Now I can buy a soda and spend money.	Capability Commerce_buy Money Temporal_collocation

Table 3.2: Example input record

Sentence	m-Frame Types	m-Frame
Now I can buy a soda and spend money.	FrameNet Frame names	Capability Commerce_buy Money Temporal_collocation
Now I can buy a soda and spend money.	Nouns	I soda money
Now I can buy a soda and spend money.	Mixed	Capability Commerce_buy Money Temporal_collocation I soda money

Table 3.3: Example m-Frames

The input files are sectioned using 10-fold cross validation and the training sample is 10% of the total input records. For each round of the experiment, the L-LDA algorithm trains a topic model. This forms the gold standard. The topic modelling toolbox is used again, to infer against the held-out 90% and the result is a probability distribution over topics and sentences. The comparison of the gold standard with the predicted (or *infered* as described by the Stanford toolset) is performed, and the resulting performance figures were obtained.

3.3.3 Results

The aim of this experiment was to determine whether the Mental Spaces model, with the Latent Dirichlet Allocation algorithm overlaid, can be used as a classification tool. Our intuition was that the different m-Frames would have a different result - some m-Frames are intuitively more influential than others. We ran various iterations of the model with various parameters that were altered at each iteration. For each m-Frame type, and by changing the values for the Term Smoothing, Probability Cut-Off and Topic Cut-Off values we achieve different results. The results for each analysis are shown in Tables 3.4 and 3.5. Where parameters achieved the same performance results, we have indicated the range, e.g. for Probability Cut-Off, 0.05 - 0.1 indicates that all probabilities in this range achieve the same result.

m-Frame	L-LDA Term Smoothing	Probability cut-off	Topic cut-off	Macro F1
AllInheritedFrames	0.01	0.05 - 0.1	29	0.23
AllInheritingFrames	0.01	0.05 - 0.07	29	0.15
Causative	0.1	0.12 - 0.14	10	0.11
CausativeStatice	0.2	0.1	10	0.08
Earlier	0.2	0.13 - 0.14	10	0.13
Frame	0.01	0.05 - 0.07	29	0.13
FrameElements	0.01	0.05	29	0.05
HasSubFrame	0.2	0.1 - 0.11	10	0.20
Inchoative	0.5	0.12	10 - 29	0.08
InchoativeStative	0.5	0.13	10 - 29	0.10
InheritsFrom	0.01	0.1 - 0.12	23	0.20
IsInheritedBy	0.01	0.09	28	0.19
Later	0.2	0.12	10	0.16
Manually-annotated	0.01	0.11	10	0.09
Neutral	0.1	0.05 - 0.14	10	0.16
Perspectivized	0.01	0.05 - 0.07	10 - 29	0.15
POS	0.01	0.05 - 0.12	20	0.44
POSGroup	0.01	0.05 - 0.06	10 - 29	0.89
Referred	0.2	0.14	10	0.16
Referring	0.5	0.14	10	0.11
SubFrameOf	0.1	0.08	10	0.14
UsedBy	0.01	0.09 - 0.1	29	0.17
Uses	0.01	0.1 - 0.11	26	0.18

Table 3.4: Text Classification Experiment Results by best m-Frames - Macro F1 scores

m-Frame	L-LDA Term Smoothing	Probability cut-off	Topic cut-off	Micro F1
AllInheritedFrames	0.01	0.05 - 0.1	18	0.68
AllInheritingFrames	0.1	0.08 - 0.08	29	0.35
Causative	0.5	0.12	10	0.15
CausativeStatice	0.1	0.05 - 0.14	10	0.09
Earlier	0.5	0.12 - 0.12	10	0.15
Frame	0.2	0.09 - 0.1	22	0.47
FrameElements	0.01	0.05	29	0.42
HasSubFrame	0.5	0.14	10	0.13
Inchoative	0.5	0.12	10 - 29	0.083
InchoativeStative	0.5	0.14	10 - 29	0.11
InheritsFrom	0.1	0.09 - 0.13	13	0.61
IsInheritedBy	0.01	0.05 - 0.1	25	0.43
Later	0.2	0.12	10	0.20
Manually-annotated	0.2	0.08 - 0.09	10	0.23
Neutral	0.01	0.05 - 0.08	10	0.31
Perspectivized	0.01	0.05 - 0.07	10 - 29	0.19
POS	0.5	0.05 - 0.14	13	0.74
POSGroup	0.01	0.05 - 0.06	10 - 29	0.90
Referred	0.5	0.11	10	0.19
Referring	0.5	0.13 - 0.14	10	0.13
SubFrameOf	0.2	0.12	10	0.20
UsedBy	0.1	0.11 - 0.12	27	0.47
Uses	0.2	0.11 - 0.12	15	0.57

Table 3.5: Text Classification Experiment Results by best m-Frames - Micro F1 scores

In total we executed over 180,000 iterations for this experiment. We present a subset of the results - the best values of Micro and Micro F1. There is a wide variation of results from this experimental subset, ranging from 0.05 to 0.89 for Macro F1 and 0.09 to 0.9 for Micro F1. We present these as a chart in Figure 3.2

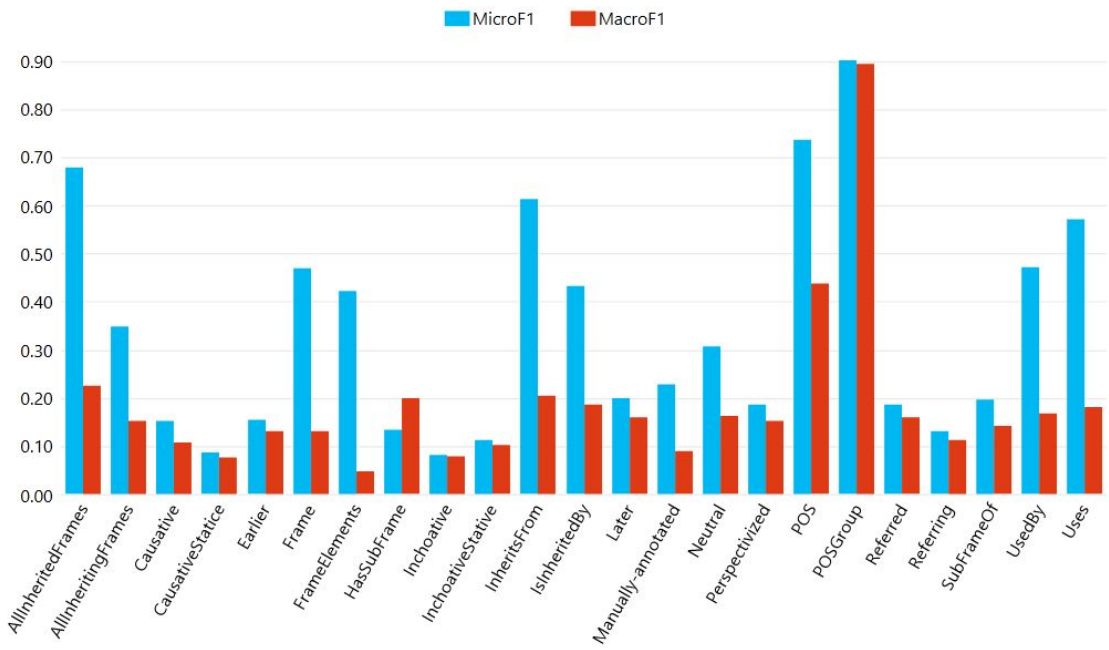

Figure 3.2: Micro F1 and Macro F1 scores by m-Frame analysis

By far the best performing m-Frame, for both Macro and Micro F1, is the POSGroup m-Frame. This is data related to the Stanford CoreNLP parsing algorithm that takes input words and assigns a part of speech to them, e.g. run -> Verb etc. All the other m-Frames score relatively low for Macro F1, with the exception of POS. The difference between POS and POSGroup could be significant. POSGroup is a grouping of more specific tags. There is not a one-to-one correspondence between them, e.g. for the word "set", the POS Group could be "Verb" and the POS could be "VBD" (Verb, past-tense). The Stanford CoreNLP parser uses the Penn Treebank for tag categories. In this way we see that the grouping performs better than the lower level POS tags.

When we look at the best performing m-Frame that includes FrameNet evoked entities, the All Inherited Frames and Inherits From perform better than 0.6 (Micro F1). There is seemingly

no correlation between the parameters that were varied in the experiment (Term Smoothing, Probability Cut-Off and Topic Cut-Off) and the performance. The AllInheritedFrames Micro F1 score of 0.68 was obtained with Term Smoothing of 0.01, all Probability Cut-Offs between 0.05 and 0.1, and Topic Cut-Off of 18. Compared with the highest Micro F1 score for the InheritsFrom m-Frame the parameters were quite different: the Term Smoothing used was 0.1, all Probability Cut-Offs between 0.09 and 0.13, and Topic Cut-Off of 13. This indicates that the parameters don't, in themselves, influence the output, but this is discussed further in later chapters.

These results compare favourably with alternative methods or text classification using semantic databases, see [Moldovan et al., 2004] for a Support Vector Machine model using FrameNet. The best results in our model (F1 of 0.9) are in advance of alternatives we have seen.

3.4 Experiment 2: Semantic Similarity

Measuring the similarity of sets of linguistic units has uses in many differing NLP tasks, such as Textual Entailment, Word Sense Disambiguation, Information Extraction and Machine Translation [Agirre et al., 2009]. Semantic Similarity is focussed, specifically, on the semantic elements of the compared language structures, for example, the meaning associated with phrases. There are different approaches to analysing the similarity of word sequences, such as knowledge-based approaches and distributional approaches. If we assume that we can assign a single measurement to the complex semantic relationship between groups of words then we use this as a factor in determining the effectiveness of computer models that attempt to compute similarity.

3.4.1 Method

In this experiment, the corpera from the Semantic Similarity Task workshop hosted at the SemEval conference (from years 2012, 2013 and 2014)[4] are used as a measure of Textual Semantic Similarity. These corpora take the form of sentence pairs with associated manual score of semantic similarity, e.g. {"The dog bit the man", "The hound bit the man", 4.8}. The scores range from 0 to 5. I have used the corpora from the 2012, 2013 and 2014 tasks.

The hypothesis for our experiment is that there is a relationship between the Mental Space

[4] http://alt.qcri.org/semeval2014/task10/

Network, resulting from the two sentences, and the manual scores. By building various candidate Mental Space Networks for these sentence pairs, running a linear regression predictive machine learning algorithm over the outputs, and then calculating a set of correlations, a value for the relatedness is determined. Figure 3.3 shows a representation of the intuition behind this experiment.

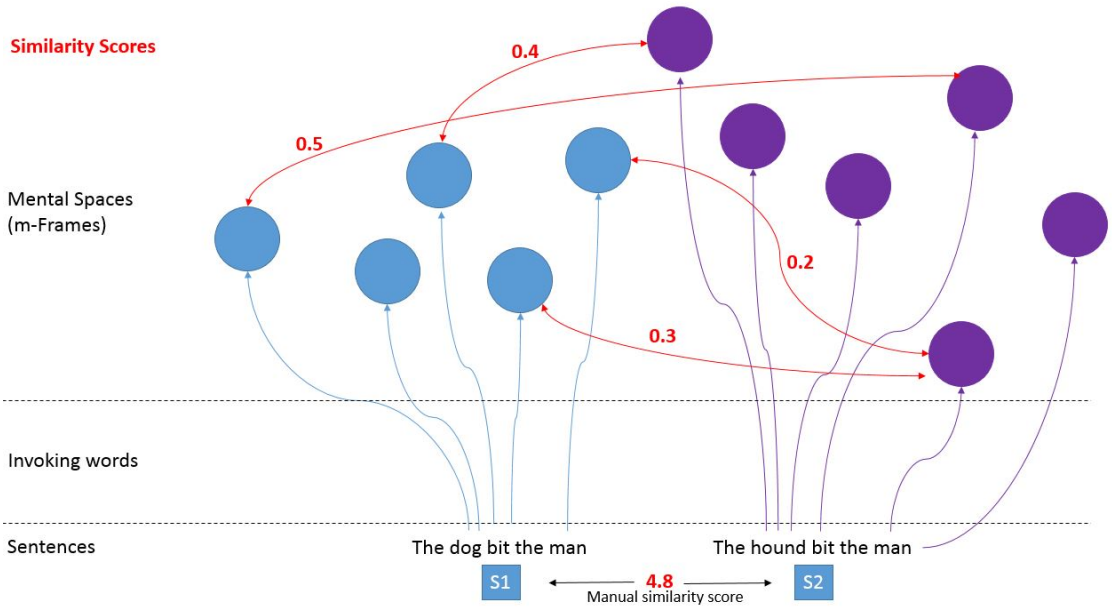

Figure 3.3: Intuition behind this experiment. Investigating the relationship between the manual similarity score and the combination of calculated similarity scores between each M-Frame.

After the inputs are parsed and pre-processed in the standard manner, we create a matrix of features for each sentence - derived from the m-Frame matrix data. In order to analyse numerically the similarity of the Mental Space Networks, we numerate the features. This is done in two phases. The first phase consists of counting the elements within each sentence or elements evoked by each sentence, e.g. if there are 3 frames evoked by the sentence then the numerical value in the 6th column of the numerical output would equal 3. Table 3.6 shows an example of the resulting m-Frame matrix with first phase counts. The purpose of the first phase is to create a measurement from which to generate a single figure that represents the overlap or similarity between sentences (we count duplicates since this is potentially a factor in the method - that duplicates could indicate importance or value). There are many ways to do this, but we settled

on those described in Table 3.7, which rely on the three further calculations performed in this phase, listed below:

- (a) numerate the number of elements in each feature for sentence 1 that match the elements in the corresponding feature in sentence 2. For example, for the feature POS, taking each element in sentence 1 in turn we compare with each element in sentence 2 and numerate through. The elements from sentence 1 that match elements in sentence 2 are numerated as NNP, NNP, VBD, TO, VB, JJ, NN, which equals 7.

- (b) number of elements from each feature in sentence 1 that are also to be found in the corresponding feature for sentence 2. For example, for the feature POS, we take each element from sentence 1 and incremenent the count where there is a single match in sentence 2, therefore since NNP, VBD, TO, VB, JJ and NN all appear in sentence 2, the total equals 6.

- (c) number of elements from each feature in sentence 2 that are also to be found in the corresponding feature for sentence 1. For example, for the feature POS, we take each element from sentence 2 and incremenent the count where there is a single match in sentence 1, therefore, since NNP, NNP, VBD, TO, VB, JJ and NN all appear in sentence 1, the total equals 7.

Feature	Sentence 1	Sentence 2	count 1	count 2	(a)	(b)	(c)
Input Sentence	Netanyahu set to call early vote	Israel's Netanyahu set to call early vote	6	7	6	6	6
POS	NNP VBD TO VB JJ NN	NNP POS NNP VBD TO VB JJ NN	6	8	7	6	7
POS Group	Noun Verb Verb Adjective Noun	Noun Noun Verb Verb Adjective Noun	5	6	11	5	6
Lemmata	Netanyahu set to call early vote	Israel 's Netanyahu set to call early vote	6	8	6	6	6
Nouns	Netanyahu vote	Israel Netanyahu vote	2	3	2	2	2
Named Entities	Netanyahu	Israel Netanyahu	1	2	1	1	1
Frames	Bail_decision Change_of_consistency Placing Cause_change_of_consistency ...	Bail_decision Change_of_consistency Placing	16	16	16	16	16
Frame Elements	Place Status Means Judge Time ...	Place Status Means Judge Time ...	156	156	532	156	156
Frame LUs	bail.n bond.n fix.v set.v order.v soften.v ...	bail.n bond.n fix.v set.v order.v soften.v	353	353	587	353	353
Is Inherited By	Dispersal Besieging Invading Labeling	Dispersal Besieging Invading Labeling	4	4	4	4	4
Perspectivized On			0	0	0	0	0
Uses	Communication Motion Simple_name Judgment_communication Communication Being_named ...	Communication Motion Simple_name Judgment_communication Communication Being_named ...	8	8	8	8	8
Used By			0	0	0	0	0
Has Subframe			0	0	0	0	0
Inchoative			0	0	0	0	0
Inchoative Stative	Cause_change_of_consistency Name_conferral	Cause_change_of_consistency Name_conferral	2	2	2	2	2
Causative	Change_of_consistency	Change_of_consistency	1	1	1	1	1
Causative Stative	Intentionally_act Event Event Objective_influence Transitive_action Event ...	Intentionally_act Event Event Objective_influence Transitive_action ...	19	19	55	19	19
Earlier	Entering_of_plea Removing	Entering_of_plea Removing	2	2	2	2	2
Inherits From	Intentionally_act Event Transitive_action Transitive_action Intentionally_affect ...	Intentionally_act Event Transitive_action Transitive_action Intentionally_affect ...	9	9	11	9	9
Later			0	0	0	0	0
Neutral	Placing_scenario Hostile_encounter Simultaneity	Placing_scenario Hostile_encounter Simultaneity	3	3	3	3	3
Referred			0	0	0	0	0
Referring	Removing Filling Time_vector	Removing Filling Time_vector	3	3	3	3	3
Subframe Of	Arraignment Cause_motion	Arraignment Cause_motion	2	2	2	2	2
Evoking words	set call early	set call early	3	3	3	3	3

Table 3.6: Example Semantic Similarity m-Frame matrix, with first phase counts

The second phase involves taking the values from the first phase and creating bespoke measures of the similarity. These are described in Table 3.7.

ID	Description
1	Number of matched entities (a, above) divided by the count of elements in sentence 1
2	Average of the number of existing examples (average of b and c, above)
3	For each feature, the absolute difference between the total number divided by the sum

Table 3.7: Similarity measurements, also called Measure IDs

Once the matrix is in place, the hypothesis is tested by attempting to determine the combination of features that achieves the best correlation to the supplied manual similarity scores. To achieve this, the output matrix of the numerical similarity values is analysed via a multivariate linear regression. Using Octave[5] to define a set of parameters and coefficients, the ideal linear relationship can be determined via a Gradient Descent algorithm. The Theta parameters can then be used as a predictive model.

The data contains 27 m-Frames for each sentence which are calculated into three different banks, one for each scoring method (measures of similarity). Each scoring method is analysed independently via the linear regression algorithm. We don't know which set of features will provide the best correlation to the supplied scores so, ideally, we would try all permutations. To compute the full set of permutations of 27 independent features of variable length is intractable, however, so we used all permutations of the following numbers of features: 1, 2, 3, 23, 24, 25 - that is, each regression analysis used a set of features from the permuted input variables as singles, doubles, triples, and groups of 23, 24 and finally up to the permutation of 25 features. Using 10-fold cross validation the full set of sentence pairs from all years' SemEval tasks was analysed (11441 records). As a predictive model, the output for each regression analysis was a set of similarity scores which could be compared with the original held-out manual score.

3.4.2 Experiment Summary

What we are trying to achieve is a strong correlation between the manual similarity scores, that are provided with the corpus, and the calculated similarity scores based on the model we develop. After pre-processing, we obtain a set of m-Frames for each sentence. These are compared and numerated for similarities in order to compute a set of values for the similiarity of the sentences (the three similarity measurement calculation methods described previously, see table 3.7). An example dataset at this stage is shown in tables 3.8, 3.9 and 3.10.

[5] `http://www.gnu.org/software/octave/`

Sentence Pair ID	m-Frame 1, Measure ID 1	m-Frame 1, Measure ID 2	m-Frame 1, Measure ID 3
1	0.9	0.4	1.73
2	2.3	0.05	0.2
3	0.53	0.04	0.11

Table 3.8: Example similarity measurements between m-Frame 1

Sentence Pair ID	m-Frame 2, Measure ID 1	m-Frame 2, Measure ID 2	m-Frame 2, Measure ID 3
1	0.3	1.1	0.02
2	1.76	0.75	0.13
3	0.18	0.4	2.16

Table 3.9: Example similarity measurements between m-Frame 2

Sentence Pair ID	m-Frame 3, Measure ID 1	m-Frame 3, Measure ID 2	m-Frame 3, Measure ID 3
1	2.1	1.44	0.3
2	0.77	0.65	1.1
3	1.02	0.92	0.81

Table 3.10: Example similarity measurements between m-Frame 3

These tables show an example of a subset of data. At this stage we have built-up a set of similarity scores for each sentence pair and for each m-Frame. In order to determine the most effective combination of m-Frames, i.e. which m-Frames form the best model at predicting similarity scores, we compute a linear regression analysis. The training and test data are a 10-fold cross validation set, formed from the similarity scores per sentence pair as described above. We did not seek to understand whether this model on its own could compete with state of the art predictive semantic similarity systems - the model would, very likely, not perform at that level. Instead we are seeking to show that this approach can be useful and worthy of further study, possibly as incoroporated into a mixed approach to solving this task. Therefore we are not attempting standard predictive model output that could be validated on those terms.. Instead, because we don't compare directly the predicted similarity scores to the supplied manual scores, we correlate them to determine the best fit. Specifically, we use Octave's **cor** function to calculate

the crosscorrelation between the prediction and actual similiarity scores.

3.4.3 Results

The aim of this experiment was to compare the m-Frames evoked by sentences that were input in pairs to the model. The intuition is that the similarity of sentences, as given in the SemEval corpora, correlates with the similarity of the m-Frame sets related to each sentence in the pair. The top results for each analysis are shown in Table 3.11.

Calculation	Number of m-Frames	Correlation
Measure ID - 2	25	44%
Measure ID - 1	23	41%
Measure ID - 1	24	41%
Measure ID - 1	25	41%
Measure ID - 1	3	38%
Measure ID - 1	2	33%
Measure ID - 1	1	29%

Table 3.11: Semantic Similarity Experiment - top results

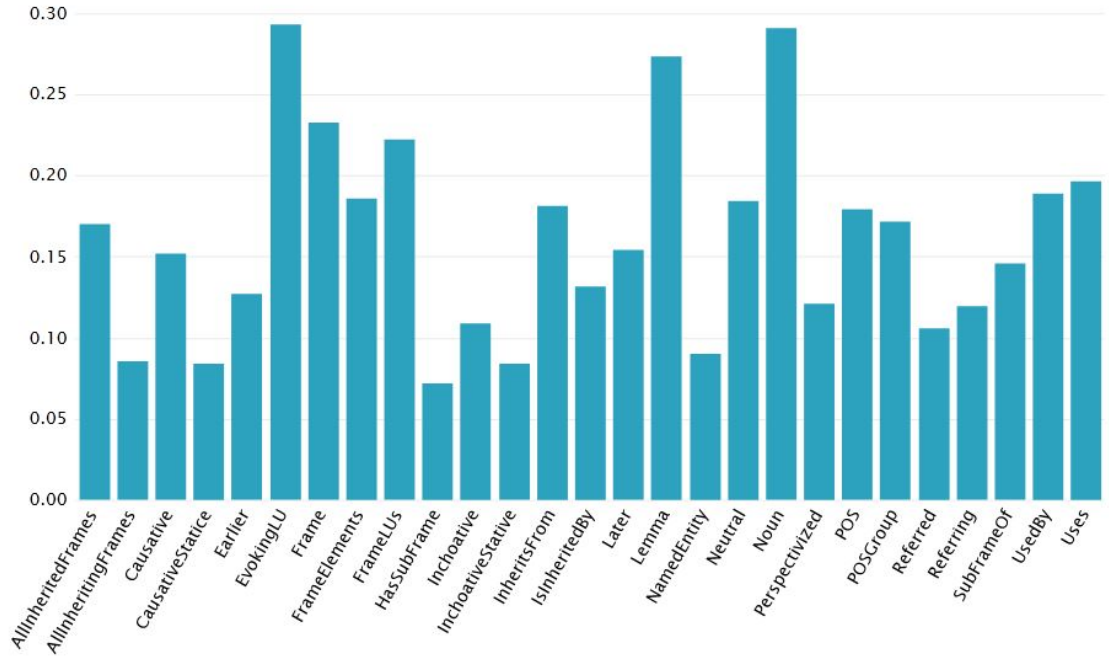

Figure 3.4: Single feature m-Frame correlation against manual similarity score

The best performing measurement is a combination of 25 m-Frames. This is intuitively sensible - the more features that the linear regression algorithm has then the more effective it will likely be. Not too far behind in terms of performance there are the triple, double and single-feature results. These diminish in effectiveness in proportion to the number of m-Frames. Again, this is intuitive, since the fewer features the regression has, the harder it would be to determine the best fit algorithm.

Looking more closely at the small feature experiments we see that the best performing m-Frames are not the ones evoked from FrameNet, see Tables 3.12, 3.13 and 3.14 for the data.

m-Frame 1	**m-Frame 2**	**m-Frame 3**	**Correlation**
POSGroup	Lemma	Noun	38%
POS	Lemma	Noun	37%
POS	Lemma	SubFrameOf	34%
POS	Lemma	Later	34%
POS	Lemma	Referring	34%
POS	Lemma	Earlier	34%
POS	POSGroup	Lemma	34%
POSGroup	Lemma	Later	34%
POS	Lemma	Inchoative	34%
POSGroup	Lemma	SubFrameOf	34%

Table 3.12: Triple feature similarity correlation - top 10

m-Frame 1	**m-Frame 2**	**Correlation**
POS	Lemma	33%
POSGroup	Lemma	33%
Noun	Named Entity	32%
Noun	Later	30%
Lemma	Noun	30%
Noun	SubFrameOf	30%
Noun	Earlier	30%
Noun	Referring	30%
POSGroup	Noun	30%
Noun	Inchoative	30%

Table 3.13: Double feature similarity correlation - top 10

m-Frame	Correlation
Evoking Lexical Unit	29%
Noun	29%
Lemma	27%
Frame	23%
Frame Lexical Units	22%
Uses	20%
UsedBy	19%
Frame Elements	19%
Neutral	18%
Inherits From	18%

Table 3.14: Single feature similarity correlation - top 10

In the above tables we can see that the majority of m-Frames are not derived from FrameNet, i.e. POS, POSGroup, Lemma, Noun and Evoking Lexical Unit.

Since we do not compute the predicted similarity score, as intended by the SemEval tasks for which the corpus was compiled, we cannot compare our results with those from the competetive task.

3.5 Experiment 3: Blending and Style

3.5.1 Method

The theory of Conceptual Integration Networks was discussed in section 2.2. We aim to explore this theory and investigate the relationship between the evoked frames in a document. In this experiment we take a corpus of text documents with associated, known styles or characteristics and calculate a measurement for the Conceptual Integration Network of each one. We call the measurement the Blending Factor. By correlating the known metadata for the text with the blending factor we attempt to validate the hypothesis that different styles of writing exhibit related patterns of conceptual integration. The intuition behind this approach is that by computing an LDA mixture model blend of the output of the Mental Space Network model, we can approximate a Conceptual Integration Network. In effect our model is much simplified and constrained.

The corpus used in this experiment is a bespoke collection of blog entries. This corpus

contains a number of blog entries from two sources: the humourist Scott Adams [6] and excerts from the Telegraph newspaper Business Section blog[7].

Each document in the corpus is a set of many sentences and each one is pre-parsed in the standard method. We develop an m-Frame matrix for each sentence, as per the previous experiments, and output a comma-separated variable file (csv) for each document (collection of sentences). In order to approximate the Blending that occurs between the many m-Frames we use a standard LDA analysis rather than Labeled LDA. In this experiment we run the algorithm over each of the m-Frames instead of the original sentence.

The output of the LDA algorithm is a probability distribution over m-Frames and topics which indicates the probability that a particular m-Frame set (which is a representation of the Mental Space Network evoked by the initial sentence) is related to a particular characteristic of the text. Figure 3.5 shows a representation of the relationship between sentences and topics.

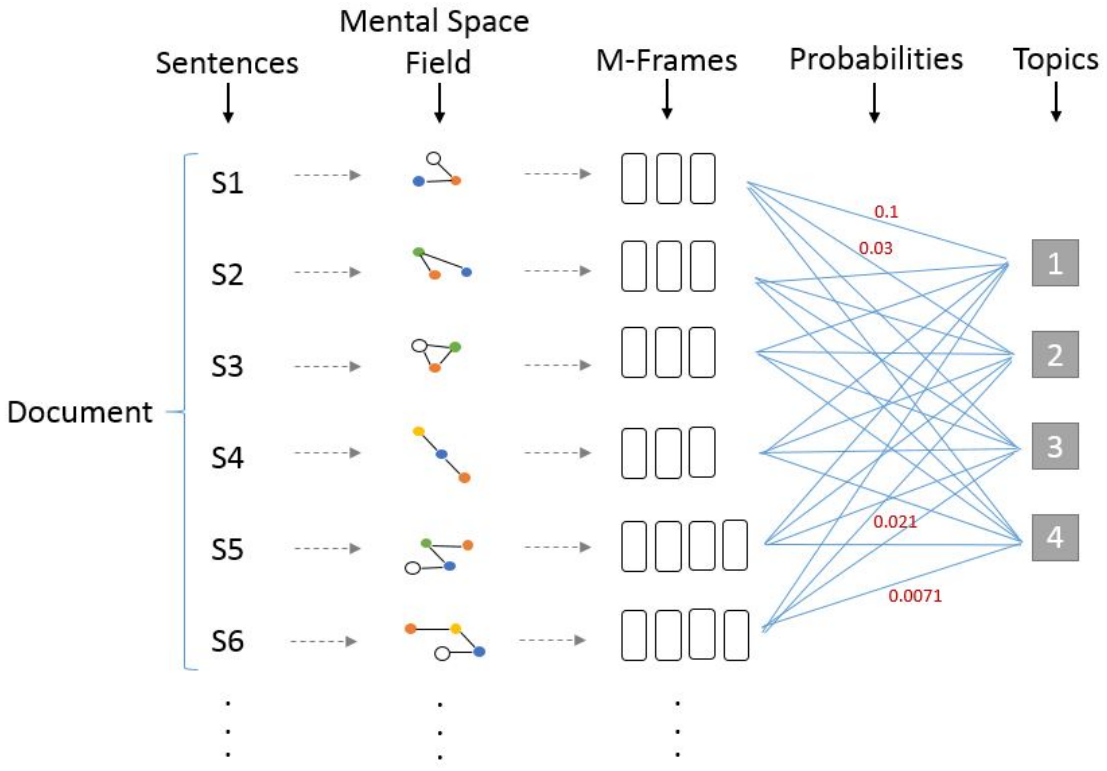

Figure 3.5: Example of sentence, m-Frame, and probability relationships to Topics

[6] http://www.dilbert.com/blog/
[7] http://blogs.telegraph.co.uk/finance/

We calculate a factor of the relatedness between m-Frames by assuming a transitive relationship over the m-Frame-topic probability distribution: all m-Frames are probabilistically related to topics and all topics are related to m-Frames by similar distributions - therefore, we assume that m-Frames are related to one another by their relatedness to the same topics. By examining the probability values and measuring the differences, we can calculate a relationship factor between m-Frames. Taking each m-Frame in turn, for each document, we calculate the difference in the probabilities between each sentence-derived m-Frame in relation to each LDA Topic. For example, looking at the m-Frame for Perspectivized On, for a document with five sentences, we would see that there are probabilities for the five sentences that relate each one to the set of LDA Topics (Table 3.15). For this set of five sentences there are ten unique permutations that do not include the sentence relating to itself - see Table 3.16 for an example of this calculation.

Document	m-Frame	Sentence	Topic 1	Topic 2	Topic 3	Topic 4	...	Sum
Telegraph Blog A	Perspectivized On	1	0.01	0.05	0.04	0.23	...	1
Telegraph Blog A	Perspectivized On	2	0.03	0.097	0.07	0.15	...	1
Telegraph Blog A	Perspectivized On	3	0.01	0.01214	0.1	0.11	...	1
Telegraph Blog A	Perspectivized On	4	0.08	0.075	0.01	0.08	...	1
Telegraph Blog A	Perspectivized On	5	0.073	0.0134	0.18	0.13	...	1

Table 3.15: m-Frame topic probability example

Sentence (X)	Sentence (Y)	Topic 1 X	Topic 1 Y	Absolute Difference	...	Sum across all topics
1	2	0.01	0.03	0.02	...	0.1
1	3	0.01	0.01	0.0	...	0.04
1	4	0.01	0.08	0.07	...	0.071
1	5	0.01	0.073	0.063	...	0.3
2	3	0.03	0.01	0.2	...	0.74
2	4	0.03	0.08	0.05	...	0.61
2	5	0.03	0.073	0.043	...	0.09
3	4	0.01	0.08	0.07	...	1.4
3	5	0.01	0.073	0.063	...	0.78
4	5	0.08	0.073	0.007	...	1.22
					Total	5.351
					Reciprocal	0.1869

Table 3.16: m-Frame topic probability permutation calculation

Calculating the difference between all the probabilities of sentence m-Frames gives us a number for how dissimilar the sentences are, in terms of their m-Frame sets. What we want, however, is a measure of the similarity between sentences. We take the reciprocal of the sum of the differences for all the sentences for the document. This measure is not a probability of relationship, but simply a relative factor that we use to calculate the next step. What the number represents is the similarity of the m-Frames evoked by all the sentences in the document. This is a crude method, as we will discuss in the conclusion, however when we correlate these numbers with the given styles of document we see the results described in the next section.

3.5.2 Experiment Summary

In this experiment we take text documents consisting of blogs from two known styles - humorous and business-related. These are pre-processed as per the normal model, in order to generate a set of m-Frames for each document. These are further processed via a standard LDA algorithm so that we obtain a probablility distribution for each document that relates each sentence to a set of topics. Using an assumption that there is a transitive relationship between sentences, e.g. sentence 1 is related to sentence 2 by virtue of the fact that they are both related to topic X by probabilities A and B, we further obtain a set of intra-document relationship values. To calculate an overall figure for the amount that sentences are similar across the document, we work out the differences between each probability, sum them and take the reciprocal value, which we call the Blending Factor.

Next we use a Mann-Whitney U test to determine the significance of the difference between the Blending Factor of each document and the known style.

3.5.3 Results

Our assumption is that humorous articles will show a different Blending pattern from business-related blogs.

This experiment's aim was to construct a mental space network m-Frame matrix for a given set of documents and, overlaying an LDA analysis, to approximate a Conceptual Integration Network. The probability distribution that results from this analysis is then used and, by way of a bespoke differential algorithm, we calculate the intra-similarity of sentences across each document. These similarity scores (Blending Factors) are then analysed with a Mann-Whitney U test. The intuition is that the amount of similarity across the document's m-Frame matrix is related to the style or theme of the document.

A drawback of our bespoke corpus is that the document counts were relatively low - 10 documents for each style - humorous and business. Due to the sparsity of the results, not all m-Frames could be calculated. Where an output from the model was available however, a Mann-Whitney U test was undertaken. This algorithm is a non parametric null hypothesis test that calculates the difference between two classes of results, with the null hypothesis being that there

is no difference between the medians of the result set. The results from this experiment are shown in table 3.17. What we are looking for is a p-value below 0.05, which would indicate that the distributions of the two groups differed significantly.

m-Frame	**n1**	**n2**	**U**	**p-value**
Inherits From	10	8	14	0.01
All Inherited Frames	9	9	22	0.057
Used By	9	8	19	0.057
Uses	8	6	15	0.14
All Inheriting Frames	7	6	13	0.15
Frame Lexical Units	9	9	30	0.19
Frame	10	10	38	0.2
Is Inherited By	7	5	17	0.31
Lemma	10	9	39	0.33
Frame Elements	10	9	44	0.48

Table 3.17: Mann Whitney p-values between Blending factor and Document type

The results show that, in keeping with the semantic similarity experiment, the correlation is not a particularly strong one. Only one m-Frame exhibited a strong difference between document types with two others close to being significant. In contrast to the semantic similarity experiment, however, the highest score in this set of results is a FrameNet-related element (Inherits From).

We recognise that the Blending Factor is a single score for the whole document. We derived it this way in order to correlate against other factors that are at the document level. When we construct the Blending Factor, however, it's clear that there is a pattern across the whole document that ought to correlate with the pattern of conceptual integration. For example, Figure 3.6 shows a visual representation of this landscape. The 17 sentences in this document, analysed in this diagram are transitively related to each other in irregular proportions, indicated by the height of the spikes in the chart. This feature of the analysis was not investigated further, but highlights a potential further investigation into blending at a sub-document level.

We believe this approach to be novel, such that comparible performance measurements from existing systems are not available.

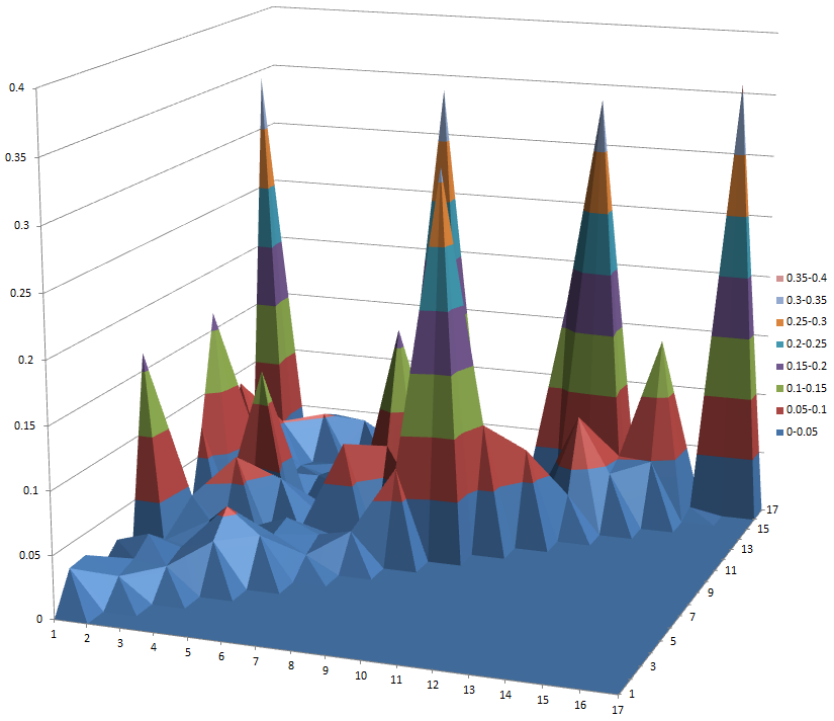

Figure 3.6: Blending Landscape - a topographical correlate to Conceptual Integration

3.6 Summary of experiments

The aim of these experiments was to create a basic model of mental spaces and conceptual integration upon which to run a number of experiments. The results of these experiments was quite mixed. Some very good results were obtained in experiment one (Classification Micro F1 of 0.9), however generally the results were inconclusive. We discuss and interpret the nature of the results in the next section.

Chapter 4

Discussion

In this research we take the problem of computing the semantic and pragmatic content of natural language. We use ideas from Cognitive Linguistics, specifically Gilles Fauconnier's and Mark Turner's theories of Mental Spaces and Conceptual Integration and Charles Fillmore's theory of Frame Semantics. These theories create frameworks to encapsulate the complex meaning in scenes and situations and also to model the way our brains process information into a meaningful set of interrelated entities. We began by using these theories to model their frameworks in a computer system, taking text as our data and distributing it across the sets of m-Frames we build making use of FrameNet's semantic network to drive the distribution. We construct a model of a Mental Spaces Network and also execute a Latent Dirichlet Allocation algorithm over the model in order to mimic a kind of Conceptual Integration Network (Blending).

We chose three experiments to test the validity of the model and to apply these techniques to known problems:

- Classification task, using the manually-tagged data supplied with FrameNet
- Semantic Similarity task, using data from the SemEval workshop
- Blending and Writing Style

We will discuss each experiment in turn and then a combined analysis at the end.

4.1 Classification Task

The results from this experiment were surprising. We make note of the differences between the types of m-Frames, particularly in reference to Mental Spaces theory. Our model collects entities into m-Frames based on what we provide via the pre-processing steps. Early in the project we chose not to restrict our model to only outputs from FrameNet evocation. For example, we decided to construct m-Frames of all the Nouns from the sentence, or all the Named Entities. Further, we included m-Frames composed of the Part of Speech tags for all the words in the sentence, see Table 3.6 which shows an example that includes Nouns, POS, POS Group etc. We included these non Frame-Net-evoked elements as a comparison with the frame-related information from FrameNet. We didn't expect these to outperform them! This is not difficult to understand when we consider the sparsity of some of the FrameNet coverage.

The best measures come out at 0.9 (Micro F1). This is a very good result and one that needs further investigation as a potential new method for classification. It is, however, not a measurement that has come from a mental spaces m-Frame and therefore does not validate the purpose of this research that derives from using Frame Semantics. The m-Frame related to our initial theory that scores highest is the All Inheriting Frames (0.68). This is a good score and one that could initiate further investigation in its own right. The corpus used in this experiment is small and, in order to further investigate these results, it would be necessary to extend the size of the dataset.

4.2 Semantic Similarity task

The results obtained in this experiment were disappointing. Overall, the best correlation obtained was 44%. This is not a level that can show a successful experiment. The focus in this experiment was to evaluate the performance of the Mental Spaces approach, so we would expect that the performance might not be comparable to other research in this area that attempts to obtain the best performance across all techniques. "The best performance is achieved using a method that combines several similarity metrics into one" [Mihalcea et al., 2006]. So even though, taken in isolation, the best correlations are uninspiring, we think that the work deserves further

investigation as an approach to be combined with others in a common goal, namely accurate semantic similarity scoring.

4.3 Blending and Writing Style

This experiment, again, had results that were not as positive as had been hoped. We were encouraged to see one m-Frame that showed a p-value < 0.05, indicating a clear difference between the medians of the two data sets (humorous and business blogs). We also note that two other m-frames had p-values of 0.057 which indicates a strong difference. This is an encouraging correlation and leads us to assert that further investigation is warranted.

4.4 Analysis

Overall, the results from the experiments indicate that there is potential in these techniques, but that much further work would need to go into improving the outputs. The best results were obtained not by m-Frames evoked from FrameNet, but by the elements derived from relatively simple parsing. For example in the first experiment the Micro F1 and Macro F1 scores for POS and POSGroup were significantly beyond those for all other m-Frames. Similarly in the second experiment (Semantic Similarity), the best results came from the POS, POSGroup, Lemma and Noun m-Frames.

We explain this general trend in the results by a number of factors, such as the sparsity of the m-Frame data. For example, a sentence of 5 words will always have POS, POSGroup, Lemma and noun entity values populated in the m-Frame matrix. It is dependent on FrameNet as to whether the other m-Frames are populated, for example there may be nothing at all in the Causative or Referring m-Frames. This is a significant problem that is made worse by the relatively small corpora used. If we had larger corpora then this issue would be potentially mitigated to some extent, since the aggregation of many more sentences would give a greater amount of data. Statistical measurements abhor a lack of data and so, in the absence of values for some of these m-Frames, we would expect that the results would be mixed.

We recognise a crucial next step in this research is to validate the results obtained. This

would be achieved by undertaking statistical significance tests. The relatively small size of the corpora used, as well as the potentially low coverage of FrameNet could lead to doubts about the validity of the outcome. Certainly, further work is required and would likely take the form of a null hypothesis analysis.

The proposal we make to remedy this issue is to segment the mental space approximations differently. In this initial model we take the simplest approach and create bags-of-words sets functionally delineated, for example all the Nouns are combined into a single set for each sentence and all the other m-Frames (POS, Referring etc) are similarly bounded by their function rather than in relation to the semantic purpose in the sentence. This is not an approach that is in keeping with the original Mental Spaces theory in which mental spaces are diverse in content and semantically related. We also note again that we have not included relationships between mental spaces nor mental space types into the model. This is something that would potentially add more data to the model.

An improved model would create a set of semantically delineated entities. For example, a set containing a noun and its evoked Frames and Frame Elements would be a different method of creating the mental space approximations, and one that is more directly related to the theory.

A number of assumptions have been made in this research. For example, we overlay a Latent Dirichlet Allocation statistical analysis over the basic model that we hope approximates Blending. Some of the results in this project would indicate that there is some validity to this assumption, however this needs further investigation. The theory of Conceptual Integration (Blending) is not simply a "hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics" [Blei et al., 2003], but a conceptual network that includes highly complex semantic entities and relationships. The brain doesn't "value" the co-occurence of words as much as our LDA model. We have produced a two-step approximation to Blending - a semantic network model using FrameNet, overlaid with a probabilistic mixture model (LDA).

Chapter 5

Conclusion & Future Work

In the previous chapter we discussed the various experiments undertaken in this research and analysed the results. In this final chapter, we take this analysis and contextualise it based on the assumptions and related theories. We also look at the future directions that are highlighted as potential research areas.

We described earlier in this report the desire to utilise pragmatic theories, e.g. context and knowledge. We see Pragmatics as crucially important for the development of computer systems that can work with natural language in a capable way that is useful for humans. This is a key goal for Artificial Intelligence research. The goal of this report is to draw a connection between the models we built - that use FrameNet and partitioning mechanisms to address problems in Computational Linguistics - and the larger goal of seemingly intelligent language systems.

What became evident early on in the project, is that the computing power necessary to compute the model would be significant. Even the simplest of models would take hours to compute some of the complex experiments. This was due to the amount of data manipulation required, especially the LDA calculations.

We were also hampered by the lack of a large corpus for this research. We had a number of manually annotated corpora, but they were all fairly limited in their volume. This was a drawback to the research. We would like to run the same experiments on larger corpora for comparison.

5.1 Extension of the model

The computer model we constructed is limited in a number of ways. To create a useful model of human cognition is a highly complex and difficult task. We chose to model the theories of Mental Spaces and Frame Semantics and we managed to implement a small subset of these theories. The Cognitive Linguistic theory of Fauconnier is large and detailed and we only scratched the surface of the potential for a computer model. For example, we did not investigate the idea that across a text there exists coreference and dependency chains (both readily available via parsing tools such as the Stanford CoreNLP Parser). We explored this in an early phase of the research, but it became difficult to incorporate with the various other m-Frames that we finally put in place.

We would have liked to perform more machine learning algorithms in the experiments we undertook. For example in the semantic similarity experiment we created a linear regression model. The next logical step would have been to investigate the same data via a neural network model. This kind of statistical model would match better the dynamic system aims set out at the start of this report.

Our research does not model mental space theory further than the very rudimentary. The reasons are that the theory is complex and the computer processing necessary becomes a difficult problem to overcome. In future research we would expect that differentiating the mental space types and recording the relationships and entity types in a more comprehensive and meaningful manner would be beneficial. This would mimic the theory's proposed real-world mental model and therefore have greater potential for success.

Our model makes no assumptions about the semantic relationships across the text in the order in which they appear throughout the text. Almost all texts have a semantic order in which it is assumed they will be processed. For English this is almost always that words appearing in the top left of a page are processed before all others and that by working one's way down (from top to bottom) the page the meaning will gradually unfold. This is not always the case in other languages, for example, Arabic texts move from right to left to unfold meaning. A drawback of our model is that it doesn't take into account any gradual increases in conceptual understanding as one reads through a text. This was a problem that we recognised and would have liked to address. For example the way that concepts are introduced and manipulated across a document

can be modelled in a more interesting way than our experiment was able to.

Further to the idea that the location of concept-evocation in a text is important, the scope of the frame analysis and overlaid statistical modelling is an area that we would like to investigate. We have used the sentence as our main grouping of meaning. This is not realistic in terms of how humans understand language. It makes the model simpler since we can define segmentation in a more ordered fashion, however for future work we would expect to see the scope of analysis change to phrases or clauses within sentences. This would enable the model to become more fine-grained with respect to the mental space evocation and segmentation.

Three factors that have been paramatised in the model and the impact on results described to some extent, are the LDA Term Smoothing parameter, the probability cut-off and the topic cut-off. The values chosen for these parameters have not seemingly made an impact on the results, however it would be necessary to further investigate their impact on the data in order to rule out the importance they may have. The Term Smoothing parameter caters for the common problem of "unseen" words appearing in documents, e.g. for a test text there may be words that have not appeared at all in the training set which can cause statistical problems ; smoothing attempts to work around this by "assigning positive probability to all vocabulary items whether or not they are observed in the training set" [Blei et al., 2003]. The probability cut-off and topic cut-off ate variables that affect the balance between Precision and Recall and therefore can be varied to hone the output. Our model is, in a way, a system that creates a large matrix of relationships between all the entities in the domain. This becomes unwieldy and not useful unless there is a way to hone the output. The probability and topic cut-offs provide that mechanism.

5.2 Limitations of FrameNet frames

Without the FrameNet database this project would not have been possible. It is a very important resource for computational semantic models - "Knowledge of semantic structure is essential for language understanding" [Palmer and Sporleder, 2010]. It is, however, limited in its coverage. This is not a criticism of FrameNet since it is growing and evolving all the time and the coverage is already large. At the moment however, it is recognised as being somewhat lacking in coverage when text-based analyses are undertaken [Palmer and Sporleder, 2010]. This is a problem for

our research since not only do we rely on as many evoked frames as possible being recognised, but also that the inter frame relationships are fully covered.

Our choices of the entities that would populate m-Frames is something that we recognise as being a variable for future work. We decided on a set of m-Frames that derive from parsed elements in a sentence, e.g. Nouns, and also m-Frames consisting of frame-to-frame relationships. The intention was to mimic a mental space network and we feel that this was achieved. It is, however, a huge task, both in terms of the analysis required and the data necessary and we believe that creating larger and more complex m-Frames that further use the FrameNet database would be a valuable avenue to explore. We expect that the use of other lexical resources would improve the model too. For example WordNet for synonyms and the population of context and/or background information via knowledge bases such as FreeBase[1] or DBPedia [2].

We also realised, while carrying out this project, that the manually-annotated corpus accompanying the FrameNet database is more detailed than originally understood. Each sentence is annotated with intra-sentence relationships between words, semantic roles are noted, and frame evocation is intra-sentence located. This information could be used to drive an analysis that is more fine-grained than ours.

The FrameNet project has been successful in its purpose. The practical benefits of a semantic database such as FrameNet have been used to generate similar projects in other languages. There are now projects working on German, Chinese, Brazilian Portuguese, Spanish, Japanses and Swedish versions of FrameNet. There are also corpora that can be used to develop experiments and computer models in languages other than English [Burchardt et al., 2006]. We would like to see the same ideas as developed in our report transferred and extended into different languages.

5.3 Conclusion

To conclude this report we say that the aims set out at the beginning have been achieved. We have designed, created and explored a computer model of Mental Spaces and Frame Semantics. We have designed and carried out a number of experiments. The results of those experiments are mixed, however, but we believe that they indicate potential for future work. The goal is highly

[1] `https://www.freebase.com/`
[2] `http://dbpedia.org/About`

ambitious and we were never under the illusion that the model could achieve a lot in a year's research. We are pleased in some areas and disappointed in others. We are hopeful that some useful applications could come from this limited achievement.

The potential applications that models of the kind generated in our research can be put towards are manifold. From machine translation to text categorisation, from semantic similarity to plagiarism detection, the kind of models we create, coupled with machine learning algorithms and Bayesian models, can be very powerful mechanisms for computational linguistics research.

Bibliography

E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics, 2009.

R. D. Beer. 6 dynamical systems and embedded cognition. *The Cambridge handbook of artificial intelligence*, page 128, 2014.

J. L. Bermúdez. *Cognitive science: An introduction to the science of the mind.* Cambridge University Press, 2014.

D. Blei, L. Carin, and D. Dunson. Probabilistic topic models. *Signal Processing Magazine, IEEE*, 27(6):55–65, 2010.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 2006.

N. Chomsky. Language and nature. *Mind*, pages 1–61, 1995.

G. Fauconnier. *Mental spaces: Aspects of meaning construction in natural language.* Cambridge University Press, 1994.

G. Fauconnier. Pragmatics and cognitive linguistics. In L. Horn and G. Ward, editors, *Handbook of Pragmatics.* Blackwell, 2004.

G. Fauconnier and M. Turner. *The way we think: conceptual blending and the mind's hidden complexities.* 2002.

J. A. Feldman. Language understanding and unified cognitive science. In *Cognitive Informatics, 6th IEEE International Conference on*, pages 1–1. IEEE, 2007.

C. Fillmore. Frame semantics. In D. Geeraerts, editor, *Cognitive Linguistics: Basic Readings.* Mouton de Gruyter, 2006.

J. Goguen and D. F. Harrell. Style as a choice of blending principles. In J. J. Shlomo Argamon, Shlomo Dubnov, editor, *Style and Meaning in Language, Art Music and Design.* AAAI Press, 2004.

T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544, 2004.

F. Hamm. Frame semantics, Nov 2007. URL `http://www.uni-stuttgart.de/linguistik/sfb732/files/hamm_framesemantics.pdf`.

D. F. Harrell. Shades of computational evocation and meaning: The griot system and improvisational poetry generation. In *Sixth Digital Arts and Culture Conference*, pages 133–143, 2005.

J. Hintikka. Knowledge and belief. 1962.

S. A. Kripke. Semantical analysis of modal logic i normal modal propositional calculi. *Mathematical Logic Quarterly*, 9(5-6):67–96, 1963.

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing.* MIT press, 1999.

R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

D. Moldovan, R. Gîrju, M. Olteanu, and O. Fortu. SVM classification of framenet semantic roles. In R. Mihalcea and P. Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, page 167–170, Barcelona, Spain, July 2004. Association for Computational Linguistics, Association for Computational Linguistics.

T. Oakley. Mental spaces. *Grammar, Meaning and Pragmatics, Amsterdam/Philadelphia, John Benjamins*, pages 161–178, 2009.

T. Oakley and A. Hougaard. *Mental spaces in discourse and interaction*, volume 170. John Benjamins Publishing, 2008.

A. Palmer and C. Sporleder. Evaluating framenet-style semantic parsing: the role of coverage gaps in framenet. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 928–936. Association for Computational Linguistics, 2010.

J. Prinz and L. Barsalou. Steering a course for embodied representation. In E. Dietrich and A. Markman, editors, *Cognitive dynamics: Conceptual change in humans and machines*. Psychology Press, 2014.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

J. Ruppenhofer, M. Ellsworth, M. R. Petruck, C. R. Johnson, and J. Scheffczyk. Framenet II: Extended theory and practice, 2006.

R. C. Schank and R. P. Abelson. *Scripts, plans, and knowledge*. Yale University, 1975.

F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

J. A. Thomas. *Meaning in interaction: An introduction to pragmatics*. Routledge, 2014.

T. Veale and D. O'Donoghue. Computation and blending. *Journal of Cognitive Linguistics*, pages 253–281, 2000.

H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

L. Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 4th (2009) edition, 1953.

F.-P. G. Yang, K. Bradley, M. Huq, D.-L. Wu, and D. C. Krawczyk. Contextual effects on conceptual blending in metaphors: An event-related potential study. *Journal of Neurolinguistics*, 26(2):312–326, 2013.