Saarland University

Computational Linguistics

Master Thesis

In *Master of Language Science and Technology*

March 31, 2008

Representing FrameNet in Semantic Spaces

Michael Roth

Evaluators: Prof. Dr. Manfred Pinkal
Dr. Marco Pennacchiotti

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Saarbrücken, den 31. March 2008

(Michael Roth)

Abstract

Semantic space models are computational representations of word meaning based on co-occurrence counts from large corpus data. This means that the meaning of each word is described by the contexts in which it occurs. Although these data-driven models have proven to be well suited for capturing a wide range of semantic information (such as similarity of synonyms and relevance of correlating words), the meaning aspects they cover have not been fully explored. This thesis examines the suitability of automatically built semantic space models for representing meaning in terms of *frame semantics*, an empirical semantic theory that emphasizes on the relation between language and experience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Frame semantics is an empirical semantic theory that "*emphasizes the continuities between language and experience"* (Petruck, 1996). Specifically, the frame semantic meaning of a word is characterized in terms of experience-based schematizations (*frames*). The underlying notion of this representation is the hypothesis that we know the meaning of a word through prototypical situations in which the word occurs[1]. For example, the meaning of the verb "kill" is established through the experience that this word is used to describe events involving a killer and a victim. In frame semantics, the killing event and its involved concepts are represented together in the KILLING frame. In general, a *frame* is a coherent structure of related concepts that (together) form an event, object or situation.

The Berkeley FrameNet project[2] (Baker, Fillmore, & Lowe, 1998) began roughly ten years ago with the goal of developing a hand-tagged corpus with frame-semantic annotations for "several thousand English lexical items". Though work in this project is ongoing, annotations have already been used in developing a so-called frame lexicon,

[1] Fillmore (1976), the inventor of frame semantics, argues, "A language-learning child first learns labels for whole situations, and only later learns labels for individual objects. A child might first associate the word *pencil*, for example, with (…) drawing circles; later on he becomes able to identify and label isolable parts of such an experience – the pencil, the paper, the act of drawing, etc."

2 Based at the International Computer Science Institute of the University of California, Berkeley

which describes prototypical frames, the words that are used to express them supported by annotated examples in terms of text fragments. Though still in development, the database has already been released to more than 80 research groups in more than 15 countries (Baker & Sato, 2003). Recently, FrameNet has proved to be useful in various language-related tasks. For example, it has been used as basis or training material for a number of applications including machine translation (Boas, 2002), question answering (Narayanan & Harabagiu, 2004), information retrieval (Narayanan & Mohit, 2003) and recognizing textual entailment (Burchardt & Frank, 2006).

The current version of FrameNet[3] consists of more than 825 semantic frames and 135,000 example sentences with 10,000 different lexical units[4]. A *lexical unit* (LU) is a word or a multi-word expression that evokes a frame (i.e. one meaning of the LU is a clear indicator for a certain frame). A non-ambiguous example for a LU is the word "kill" which indicates the frame KILLING. Other LUs for this frame include the verbs "murder" and "eliminate", the nouns "kill" and "suicide", and the adjectives "deadly" and "fatal".

One issue in FrameNet is the fact that all development steps are done manually. This means that annotators have to carefully read each sentence, determine the lexical unit, its associated frame and select phrases in the sentence that represent properties of the frame. Moreover, developers have to build new frames when needed because the frame lexicon is not yet exhaustive. Each of these development steps is highly time-consuming and expensive because they need to be done by experts. Though recent research (Gildea & Jurasfky, 2002; Fleischman, Kwon, & Hovy, 2003) showed that the existing annotations can be used as training data for automated labelling, the benefits of such an approach are limited. Aside from missing reliability, one disadvantage of this method is that automatic annotations can only be computed for already trained cases.

3 As of November 2007, the latest release is version 1.3 (June 2006).

4 http://framenet.icsi.berkeley.edu/index.php?option=com_content&task=view&id=40

Moreover, such a method only helps in expanding the database of example sentences. It cannot be used for expanding the list of lexical units.

The goal of this thesis is to overcome this deficiency by exploring and examining the suitability of unsupervised learning techniques to automatically build a full representation of FrameNet that allows for the classification of seen and unseen input data alike.

Previous work showed that it is possible to automatically compute a FrameNet-like representation for other languages using the existing database of FrameNet (see Chapter 2.3 for details). Specifically, Pitel (2008) computes a frame semantic resource for French by modelling a bilingual vector space model based on aligned English-French corpora. Starting from Pitel's promising results, we investigate the use of vector spaces for modelling FrameNet in more detail.

1.1. Semantic Space Approach

Based on the results of Pitel, we explore unsupervised training methods to learn vector representations of LUs and frames in a multi-dimensional space. We utilise the vector space model originally developed by Salton et al. (1975). Though the original application for this model is to represent documents in information retrieval (Klavans & Kan, 1998), it has been generalized to describe also smaller structures such as paragraphs, sentences and individual words by their co-occurring words.

Nowadays, modifications of this model have been used in a number of natural language applications (cf. Chapter 3 for details) and proved to be a good representation for different kinds of (lexical) semantic information. For example, vector space models have showed usefulness in identifying antonyms, synonyms and associations (Sahlgren, 2006; Lin, 1998), disambiguating word senses (Schütze, 1998) and clustering verb classes (Schulte im Walde, 2006).

The meaning of frames, however, is different from the meaning of a single word. While a word has some sort of a standalone meaning, a frame is characterized by the interac-

tion of its elements. Due to this complexity, the primary goal of this thesis is to examine whether semantic spaces are a suitable means of representing frame semantics. In order to verify this hypothesis, we will run experiments to compare vector similarity within frames and check whether vectors representing the same frame are more similar than others. For the comparison of vectors, we rely on measures that have recently proven useful for describing distributional similarity (Mohammad & Hirst, 2006).

Our hope is that the results of this work can support and contribute to semi-supervised FrameNet related tasks, such as defining relations between frames (e.g. if all vectors of a frame *A* are within the cluster of a frame *B*, *A* probably describes a special kind of *B*), and motivating new frames (i.e. if the vector of a sentence is not similar to any frame vector). Our intuition here is that similar situations are described with similar words, thus having a similar vector space representation.

1.2. Overview of the Thesis

The thesis is organized as follows: The next chapter will present an overview of frame semantics, of the FrameNet project and of state-of-the-art methods for its automatic expansion. In Chapter 3, we will introduce the concept of semantic spaces and briefly discuss different approaches and applications in which semantic spaces are used. Chapter 4 gives a detailed description of our approach and its implementation. Chapter 5 shows different experiments we performed to evaluate our methods. Finally, Chapter 6 summarizes our work and results, and discusses how further improvements could be achieved.

Chapter 2

Frame Semantics

Frame semantics is an empirically motivated sub-discipline of semantics that studies the combined meaning of a coherent structure of related concepts. In contrast to other semantic approaches, frame semantics provides no standalone meaning representation for single words, but only for so-called frames (however, single words can evoke a frame or be a property of the same). A *frame* typically represents the meaning of a situation, object or an event including its participants, properties and other related conceptual roles (*frame elements*). For example, the COMMERCIAL_TRANSACTION frame represents a situation that always consists of a buyer, a seller, money and goods. In addition to those *core roles*, a frame can also have optional (*non-core*) roles. In this case, these are the medium of exchange, the currency of the money and the payment rate per unit.

Frames are psychologically motivated by the fact that the understanding of linguistic expressions requires a complex knowledge of related background. A good example for this assumption is the word “widow”, whose meaning requires an understanding of concepts such as family, marriage and death. In frame semantics, each of these concepts is a frame of its own that gets evoked by the word “widow”. In general, frames can be evoked by a number of words which have a semantically related meaning. Examples for the previously mentioned COMMERCIAL_TRANSACTION frame include the words “buy” and “sell”. Most of these so-called *lexical units* (LU) are verbs and nouns

but other grammatical categories (e.g. adjectives and adverbs) are possible as well as multi-word expressions and idiomatic phrases.

The following sections give an overview of the frame semantic resource FrameNet (2.1), a discussion of examples from the FrameNet database (2.2), and a summary of previous attempts to automatically expand this database (2.3).

2.1. The FrameNet Project

The Berkeley FrameNet project is currently building a frame semantic lexicon for English that includes over 10,000 LUs (Ruppenhofer et al., 2006). For more than 6,000 of them, the lexicon contains annotated example occurrences taken from the British National Corpus[5]. Annotations are done manually and include tagging of lexical units and *frame element fillers* (*FEF*). Since lexical units in FrameNet are always pairs of word/meaning, tagging a LU consists of two tasks: 1) marking the frame-evoking expression and 2) selecting the right meaning (frame) for the given context. A FEF is a word or phrase that fulfils a specific role in the respective frame. Thus FEFs are tagged after the frame has been selected. Figure 1 illustrates the annotation process of an example sentence.

[5] A 100 million word corpus of written and spoken English, http://www.natcorp.ox.ac.uk/.

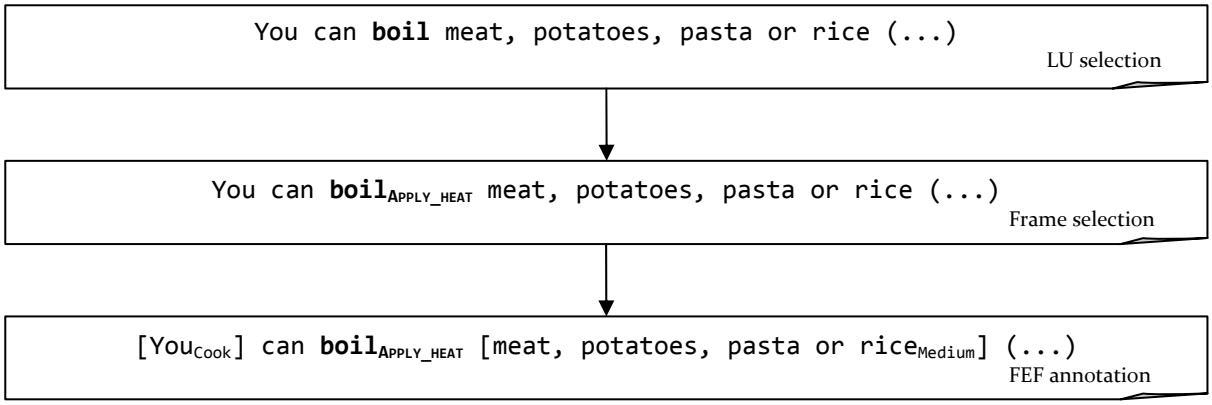

Figure 1: Annotation process in FrameNet

In FrameNet, annotation is carried out in two ways: In the lexicographic annotation mode, the annotators have to edit selected sentences that contain a particular LU. This means that the selection of LUs can be omitted, so that only the right frame and its properties have to be chosen. The aim of this mode is to collect as many different combinatorial possibilities of a word as possible. Contrary to that, annotators have to select words or phrases as LUs themselves in the full-text annotation mode. The advantage of this mode is that it allows for the discovery of frame-evoking words that are not yet considered in the FrameNet database. However, both of these annotation modes are highly time-consuming because the annotators have to carefully read each sentence and determine all frame-relevant properties.

As of November 2007, the current version of FrameNet (1.3) consists of 10,195 LUs in 795 frames, making an average of 13 lexical units per frame. However, the real number of LUs per frame varies from 0 to 179. In fact, 74 frames (9.3%) have no lexical unit at all. All LUs together divide into 39.6% nouns, 36.8% verbs, 17.1% adjectives, 5.3% multiword expressions and 1.2% other categories (prepositions, adverbs, conjunctions and interjections).

In our work, we also use the term LU when referring to a frame-evoking expression without implying a specific meaning. If a LU in this context can evoke different frames, we call it a *polysemous LU*. Currently, the FrameNet lexicon contains 1.352 polysemous LUs and 8.310 different frame-evoking expressions in total.

2.2. Discussion of Examples

To give a better understanding of the variety of possibilities in which a frame can be linguistically realized, we present a few chosen examples[6] for the word "burn". With respect to frame semantics, "burn" is a polysemous LU since it can evoke four different frames: PERCEPTION_BODY, EMOTION_HEAT, CAUSE_HARM, and EXPERIENCE_BODILY_HARM. For each of these frames, we will have a look at examples showing the syntactic and semantic behaviour of the LU.

1) PERCEPTION_BODY

According to FrameNet's definition, "burn" in this sense describes "*physical experiences that can affect virtually any part of the body. The body part affected is almost always mentioned with these words. It is typically expressed by the noun heading the external argument, and this noun is typically accompanied by a possessive determiner that refers to the possessor of the body part*".

The core roles of this frame are Experiencer and Body_part.

"... [[*his* Experiencer] *cheeks* Body_part] begin to **burn** all the way up to his scalp."
"[*The throat* Body_part] **burns** like coals of fire; ..."
"[[*His* Experiencer] *face* Body_part] **burnt** like a brand ..."

The examples show that the linguistic realizations go for the most part with the description made in the definition. This is also true for other LUs of this frame including "ache" ("*Her back* was **aching** badly"), "hurt" ("*My broken finger* **hurt** like hell") and "itch" ("*Leith's right hand* started to **itch** again").

[6] Examples are taken from the FrameNet database.

2) EXPERIENCE_BODILY_HARM

"Burn" as a LU of the frame EXPERIENCE_BODILY_HARM describes a harmful experience caused by an injury to a body part. Core roles of this frame are again Experiencer and Body_part.

"[*My sister* Experiencer]'s **burnt** [*her arm* Body_part] ..."
"[*I* Experiencer] tried it once and **burnt** [*my mouth* Body_part] ..."
"Then [*Austin Bessie* Experiencer] *(...)* **burned** [*herself* Body_part] rather badly."

In contrast to the previous meaning, the FEF for Experiencer is mentioned explicitly in the given examples. The same observation can be made for other LUs of this frame such as "break" ("*Amelie* fell and **broke** *her hip*"), "cut" ("*Stirling* **cut** *his eye* quite badly") and "strain" ("*She* **strained** *her back* at college").

3) CAUSE_HARM

In the sense of this frame, "burn" refers to a situation where an Agent hurts or kills a Victim.

"[*Up to 65 protesters* Victim] were (...) **burned** to death ..."
"[*A battered wife* Agent] *who* **burned** [*her brutal husband* Victim] to death was ..."
"Indeed, [*Henry V* Agent] used to **burn** [*them* Victim] alive."

As it can be seen in the examples, the frame elements of this frame are different from the previous two causing the dependents of the LU to be filled with words of other semantic categories. Examples with other LUs of this frame: "*She* **bashed** *the judicial scalp*", "Rodomonte saw *his father* **beating** *his mother*", and "*She* would **cut** off *her right arm".*

4) EMOTION_HEAT

The frame EMOTION_HEAT denotes a strong emotional experience. Necessary frame elements include an Emotion, an Experiencer and a Seat_of_emotion.

"... [*he* Experiencer] was heavy and **burning** [*at heart* Seat_of_emotion]
with [*his longing to ask* Emotion] ..."
"[*The desire* Emotion] **burning** [*inside* [*her* Experiencer] Seat_of_emotion] ..."
"[*The flame that* Emotion] had been **burning** [*inside* [*her* Experiencer] Seat_of_emotion] ..."

The difference between these examples and those from the previous frames is similar to the statement made for the CAUSE_HARM frame. Instead of body parts or people, the FEF are words from other semantic categories. E.g., Seat of emotion is typically filled by a word from a category that can be described as something body interior and the filler for Emotion is an abstract or metaphorical concept representing a feeling. For other LUs of this frame, the Emotion is more often filled by a feeling itself, e.g. "*The frustration that* was **boiling** inside her", "The Lemarchand woman who would be (...) **fuming** *with impatience*", and "Her chin rose as she **seethed** *with anger*".

As these examples have shown, the semantic content of linguistic realizations varies from frame to frame. In many cases using FEFs for a frame that makes perfect sense for another would produce sentences that are rather absurd, e.g.:

- *[*His cheeks* Body_part] **burned** [*himself* Experiencer] rather badly.[7]
- *[*Henry V* Agent] used to **burn** [*the throat* Body_part] like coals of fire.[8]
- *[*The desire* Emotion] [*inside* [*her* Experiencer] Seat_of_emotion] **burned** [*her brutal husband* Victim] to death.[9]

[7] Experiencer from PERCEPTION_BODY and Body_part from EXPERIENCE_BODILY_HARM

[8] Agent from CAUSE_HARM in a sentence from PERCEPTION_BODY

[9] Mixed up Frame Elements from CAUSE_HARM and EMOTION_HEAT

This fact gives us reason to believe that there is a dependency between each frame and combinatorial possibilities in how to express instances of the same. Recent research on word sense disambiguation (WSD) gives us support on this assumption: Carroll and McCarthy (2000) pointed out that *selectional preferences*, i.e. frequently co-occurring words in a grammatical relation, can give cues to noun word senses. Based on subject and object fillers, this work was extended to verbs and also showed disambiguation improvements on verb senses (McCarthy, Carroll, & Preiss, 2001). Most recently, Patrick Ye and Timothy Baldwin (2006) presented a WSD system that showed improved performance when using selectional preferences for arguments and adjuncts of verbs. We thus believe that selectional preferences could also be helpful to disambiguate LUs belonging to multiple frames.

To a lesser extent, a difference in the syntactic structure for the realizations can be seen as well. For example, it is more probable that a prepositional phrase headed by "like" is an adjunct for sentences of the frame PERCEPTION_BODY than for EXPERIENCE_BODILY_HARM:

- His face **burned** *like fire/acid/a brand/a torch*.
- ?He **burned** his arm *like* …

Previous work indicating a similar syntactic behaviour for semantically similar verbs (Levin, 1993) gives us reason to believe that this phenomenon could be used to define frame specific syntactic features. In general, we take these observations as important cues in modelling a suitable semantic space.

2.3. Automatic Approaches in Frame Semantics

Our work falls in line with other approaches to automatically building or expanding a frame semantic resource. One noticeable attempt in the FrameNet project is done on annotating semantic roles (Gildea & Jurasfky, 2002). Gildea and Jurafsky propose a trainable statistical classifier that identifies and classifies frame elements in a sentence

given its frame-evoking lexical unit. Their classifier uses conditional probabilities of syntax-semantic features extracted from example sentences. Fleischman et al. (2003) suggested improvements to this model by using a maximum entropy classifier that considers two further features: the previously assigned semantic role tags and patterns of all semantic role assignments in a sentence. Their enhanced model yields an identification precision of 73.6% and a recall of 67.9%. Including the classifier's performance, their final f-measure score is 57.6%.

Noticeable efforts have also been spent on the automatic disambiguation and expansion of lexical units. The supervised disambiguation system by Erk (2005) showed that a combination of syntax-semantic features can be used to identify frames in a sentence with a precision of 75% and 74.4% recall. The features of this system were trained on FrameNet's example sentences and consisted of co-occurring contexts including lemma and part-of-speech tags, word n-grams centered on the target word and head words of complements and adjuncts of the lexical unit. Another system, Detour (Burchardt, Erk, & Frank, 2005), aims to expand the frame lexicon by using taxonomy distances in WordNet (Fellbaum, 1998) to map unknown words to potential frames. For this task, Detour computes a set of words that are related[10] to the considered word. The most probable frame can then be determined by a heuristic that regards the number of LUs of each frame in the set of related words.

Besides work on FrameNet, recent research also proposed methods to automatically build frame semantic resources for other languages. For example, Padó and Lapata (2005) utilized FrameNet for constructing an analogous structure for French and German based on shallow parsing and automatically aligned bilingual corpora (English-French and English-German). The aligned word data is used to create a candidate list for lexical units that are then filtered by different criteria (possible alignment errors,

[10] WordNet relations considered in Detour are synonymy, hypernymy and antonomy.

polysemy and information entropy). The highest F-score obtained by evaluating the list of computed candidate LUs is 58% for German and 51% for French.

Another approach is taken by Fung & Chen (2004) who use bilingual dictionaries and the HowNet[11] ontology to transfer lexical units from English to Chinese. In a first step, bilingual resources are used to look up potential candidates. In two further steps, the HowNet category of each candidate word sense is used to compute the most probable categories per frame. The result of this method is an n-to-n mapping from LUs to HowNet concepts, which can be used as a Chinese frame lexicon. Their evaluation of this lexicon based on manually translated example sentences yielding an 82% average F-measure on lexical entry alignment.

Recent work by Pitel (2008) suggests using a bilingual vector space to build up a frame lexicon in a foreign language. In his approach, aligned English-French corpora are merged together document-wise to construct a bilingual vector space. The dimensions of this model, which contains English as well as French words, is then reduced via LSA. The idea behind this method is that the representation of the known English LUs should be similar to their unknown translation equivalents in French since both should occur within the same range of (merged) documents. When tested against a gold standard of manually annotated French, the best performance on frame target classification was 58.9% precision and 58% recall.

The results of Pitel's work indicate that words evoking the same frame occur within the same sort of contexts (across languages). Though his model only consisted of LUs that are already in FrameNet, it can be generalized to capture unknown words as well. In our approach, we evaluate the correspondence between the (monolingual) vector representation of a word and the frames evoked according to FrameNet. A positive

[11] http://www.keenage.com/html/e_index.html

outcome of this evaluation would imply the possibility of augmenting the FrameNet lexicon with unknown words based on their representation in a semantic space.

Chapter 3

Semantic Spaces

The underlying notion of semantic spaces is to represent the meaning of one word by words that can be used within its context. This approach derives from the so-called *distributional hypothesis* which states that *"the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities"* (Harris, 1968). Following this hypothesis, it is possible to model word meaning solely by counting empirical features (namely words that occur in context) extracted from text corpora.

	foot	**night**	**place**	**food**	**goods**	**insurance**	**child**
walk	610	668	504	41	30	11	498
fall	507	486	492	174	141	59	460
pay	83	281	357	246	371	657	576
buy	97	232	370	457	512	143	404

Table 1: Word-by-word matrix representing co-occurrence counts of nouns for the verbs "walk", "fall", "pay" and "buy"

Typically, the resulting model of this method is a word-by-word matrix whose rows are labelled with the words to be described and columns with the words in their contexts (cf. Table 1), or vice versa. It has to be noted, however, that this approach can also be used to describe more complex structures such as sentences, paragraphs and documents. In fact, the first implementation of a semantic vector space (Salton, Wong, &

Yang, 1975) was used in Information Retrieval and characterized documents by the words they contain.

The semantic space approach has some clear advantages over other models of word meaning (e.g. a dictionary or ontology): Firstly, the construction of meaning can simply be done by extracting context windows from a corpus, i.e. for each word to be described (*basis elements*) all neighbouring words up to an arbitrary distance (*co-occurrences)* have to be counted (post-processing such as normalization of the counts is possible but not necessary). In contrast to a dictionary that needs linguistic experts to write the respective entries, this task is usually done automatically. Consequently, the meaning of words in a special sub-domain or with respect to a certain point in time can be captured simply by choosing an appropriate corpus.

Secondly, semantic spaces allow for an easy computation of how much semantic content two terms have in common (*semantic similarity*). While dictionaries and other linguistic resources rely on manually defined relations between selected words, similarity in a vector space can be computed between all words. In a semantic space, word meaning is represented in form of a vector containing co-occurrence counts. Since vectors are constructed in such a way as to represent the same dimensions, the similarity of two words can simply be measured as the distance between the vectors. However, the disadvantage of this measure is its vagueness as it does not express any specific semantic relation.

3.1. Examples

This section contains two examples of vector space models and representations to get a better understanding of how they work. The classical approaches that are usually mentioned in literature within this context are *LSA* (Landauer & Dumai, 1997) and *HAL* (Lund, Burgess, & Atchley, 1995). Following this tradition, we will briefly discuss the outlines of both models.

1) Latent Semantic Analysis (LSA)

Originally developed for information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), LSA is a semantic space model that represents word meaning in a word-by-document matrix, i.e. the vector of each word enumerates how often it occurs in a given set of documents. In order to reduce the number of dimensions of this space, LSA uses a mathematical technique called Singular Value Decomposition (Berry, 1992). We do not discuss the details of SVD here; however it can be summarized as a way of collapsing dimensions (documents) based on their similarity (i.e. the overlapping number and selection of words in each dimension).

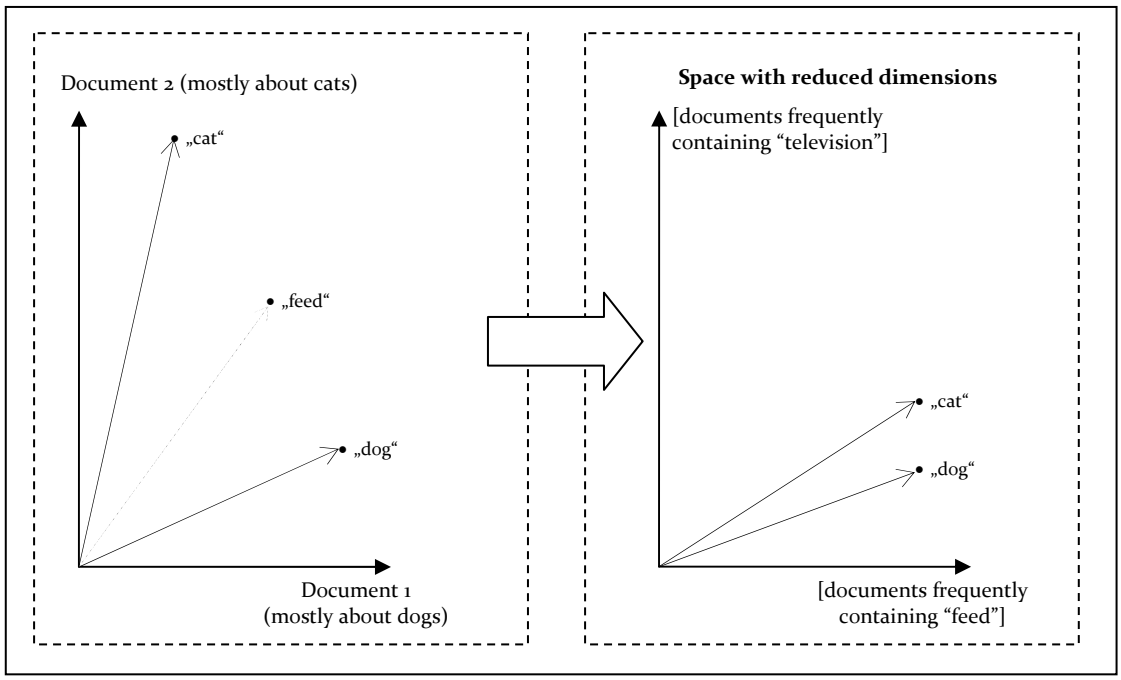

Figure 2: Simplified LSA representations before and after the application of SVD

Even though the result of this approach is a semantic space model with a relatively low number of dimensions, it can be used to preserve a wide range of information. For example, if we look at a collection of documents about **pets**, we will find that some of them may only deal with **cats** and **dogs**, while others might be about **bunnies** or **hamsters**. However, since all of them share common topics (e.g. how to **raise** pets and how to **feed** them), the use of similar vocabulary in some of them causes SVD to merge

their respective dimensions. Following that, even words referring to pets that only rarely occur within the same document can have a similar vector representation in this model. Figure 2 illustrates a simplified example[12].

2) Hyperspace Analogue to Language (HAL)

In contrast to LSA, the underlying concept of HAL is a word-by-word matrix that represents word meaning by co-occurring context words. For each word (*basis*) to be described the matrix contains one vector whose values are counts of co-occurrences with all other words. In the case of HAL, the context taken into account consists of the 10 previous and following words. Independently of the window size, words that are further away from the basis will get a lower count (cf. Table 2).

	some	**dogs**	**need**	**to**	**Be**	**fed**	**twice**	**a**	**Day**
some	0	5	4	3	2	1	0	0	0
dogs	5	0	5	4	3	2	1	0	0
need	4	5	0	5	4	3	2	1	0
to	3	4	5	0	5	4	3	2	1
be	2	3	4	5	0	5	4	3	2
fed	1	2	3	4	5	0	5	4	3
twice	0	1	2	3	4	5	0	5	4
a	0	0	1	2	3	4	5	0	5
day	0	0	0	1	2	3	4	5	0

Table 2: HAL example matrix for the sentence "Some dogs need to be fed twice a day" (with context windows of 5 words each)

Even though the word-by-word matrix uses the same labels for rows and columns, both of them are filled with different information (i.e. while one represents the context before the considered word, the other axis of the matrix represents the following context). The vector representation of a word equals the concatenated values from its re-

[12] Note that the dimensions in the reduced space no longer have concrete labels. The placeholders here are only used to give an idea of what kind of information the new dimensions represent.

spective row and column. Although a vector, whose dimensions are two times the size of the vocabulary, represents each word in this space, Lund and Burgess (1996) reported that the effects they observed only rely on the 100 to 200 most variant vector elements.

3.2. Applications

Semantic space models are tested over a variety of language related tasks, e.g. experiments that capture semantic similarity/relatedness, categorical information or semantic priming. Over the past decade, the increasing performance in such experiments made semantic space models more and more popular within the field of natural language processing. While one of the first applications of word space models can be found in information retrieval (Salton & McGill, 1983), recent research proved them to be useful in a wide range of applications including word sense discrimination (Schütze, 1998), clustering of similar verbs (Lin, 1998), text segmentation (Choi, Wiemer-Hastings, & Moore, 2001), and anaphora resolution (Poesio, Ishikawa, Schulte im Walde, & Viera, 2002).

To exemplify the use of vector spaces, we describe the word sense discrimination approach by Schütze in detail. In his work, Schütze proposed computing a high-dimensional space containing context-based vectors for each occurrence of an ambiguous word. Each vector in this space is assigned to a cluster which represents one word sense. In order to build the actual model, the context of each occurrence of the ambiguous word gets assigned to the sense cluster to which it is most similar. Since the contexts taken into account are rather small (50 words window), comparison takes place in a second order. This means that rather than comparing contexts, the vector-space representations of the contexts are compared to the representations of the contexts from the sense cluster. If the comparison yields a similarity result below a certain threshold and the number of maximum clusters is not exhausted, a new sense cluster is created. In his experiments, Schütze varied with a total number of 2 and 10 clusters,

with 10 clusters consistently out-performing. The best result achieved with this approach is a discrimination accuracy of 83.1% for naturally ambiguous words.

3.3. Discussion

As this chapter has shown, the most remarkable point about semantic space models is that they are well-suited for representing word meaning without actually incorporating any linguistic knowledge. However, this fact is not true for all semantic space models. In fact, a number of them perform a minimum of linguistic pre-processing such as lemmatization or part-of-speech tagging. Even though this preparation step can improve results, recent research also reported the possibility of decreasing results (Karlgren & Sahlgren, 2001).

One showcase for an approach that goes a step further in utilising linguistic information is the semantic space model proposed by Padó and Lapata (2007). In their work, they use a dependency-parsed corpus to build a meaning representation that relies solely on syntactically related context rather than on words in an arbitrary context window. As reported in their results, the proposed model outperforms a traditional word-based approach in experiments testing semantic priming and sense ranking.

Since traditional semantic spaces as well as the dependency-based space proved to be suitable for modelling word meaning, while at the same time showing promising results in semantic experiments, we decided to evaluate both models for building a semantic space representation of FrameNet.

Chapter 4

Frames in the Semantic Space

The fact that the meaning of a frame is more complex than word meaning makes frame modelling in the semantic space an interesting challenge. Since semantic spaces in general only represent word meaning, it is not clear whether these models are appropriate for frame-related tasks. To get around this difficulty and show the basic suitability of semantic spaces, we try to approximate a model for frames by the distribution of their respective lexical units. Even though this is an over-simplification, lexical units are the words that evoke a frame, thus being the main indicator for the same. Considering that a semantic space representation automatically comprises the context of the considered words, we expect promising LU representations to be equally adequate for frames since the contexts contain further relevant information such as frame element fillers.

The main motivation for our approach lies in the simplicity of semantic spaces. As they have proved to be well-suited for capturing phenomena such as semantic relatedness, we believe that they can be useful in the field of frame semantics as well. This is especially true seeing that instances of one frame typically refer to semantically similar objects, events or situations (cf. Chapter 2 for a detailed discussion). One possible application would be to utilize vector distances for extending the frame semantic lexicon of FrameNet. This can be done by comparing vector representations of new words to the lexical units that are already in FrameNet. An unknown word can then be assigned to the same frame as the word with which it is most similar. Another more robust

method would be to compute representations of each frame (such as by averaging over all its lexical unit vectors) and compare new words to corresponding *frame centroids*.

Yet this approach gives rise to a number of questions: Are all lexical units equally important or do some provide more frame-related information than others? Is there a way to deal with ambiguities of lexical units? Do we have to explicitly integrate other parts of the frame, e.g. frame elements? For the implementation of our method, we acknowledge these questions but cannot answer them at this point. We hope though that our evaluation in Chapter 5 will be helpful in solving some of these.

However, other questions affect the fundamental design of our model. One question that we have to answer beforehand is whether a document-based or a word-based approach is more suitable for this task. We believe that both approaches are possible but that a word-based space will work better for the following reason: Typically, a frame represents a prototypical event, object or situation. However, even though instances of such occur naturally in both context windows and documents, a document rarely focuses on one specific frame. For example, a COMMERCIAL_TRANSACTION can occur in any kind of document from business reports ("… EIE also has the right to buy the freehold from Whitbread …") to styling guides ("… Firstly, buy a good quality moisturising/conditioning mascara …")[13], which does not necessarily focus on that particular frame. This is different when only looking at the context in the same sentence, which mostly consists of relevant information in the form of frame element fillers[14]. Another consideration in this choice is the problem of sparse data. To build a reliable document-based model, there has to be a number of documents in which each LU occurs. Even though our corpus consists of 4.054 documents, they are neither normalized in length nor are they divided into any frame-specific sections. Thus, we think that a word-based model would yield more promising results.

[13] Examples are taken from http://www.natcorp.ox.ac.uk/.

[14] For example, [EIE $_{\text{Buyer}}$], [the freehold from Whitbread $_{\text{Goods}}$] in the first sentence.

4.1. Pre-Processing

Since we decided to evaluate both a classical word-based semantic space and a dependency-based space, a number of pre-processing steps are required for modelling our semantic space model. We use the probabilistic parsing system RASP (Briscoe, Carroll, & Watson, 2006) for all pre-processing steps including tokenisation, part-of-speech tagging, lemmatisation and dependency parsing. We chose RASP because of its state-of-the-art performance: 97% tagging accuracy and a lemmatisation error rate of less than 0.07% are a suitable base for reliably detecting lexical units in text because FrameNet's lexicon only contains LUs in the form of lemmas and part-of-speech tags.

RASP's output contains the pre-processed sentence followed by grammatical relations in the following format (see Figure 3 for an actual example):

```
(|NUM:LEMMA+INFLEXION_POS| |LEMMA+INFLEXION_POS| …[15])
(|RELATION| |HEAD+INFLEXION_POS| |DEPENDENT+INFLEXION_POS|)
```

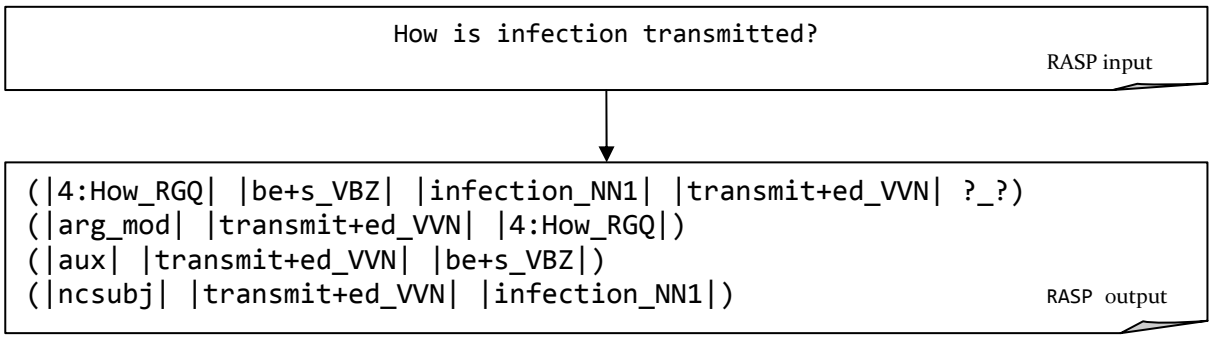

Figure 3: RASP input and output format

[15] Each line contains as many |LEMMA+INFLEXION_POS| elements as there are words in the sentence considered.

4.2. Model and Implementation

Our system goes through several processing steps to compute the final vectors in the semantic space (cf. Figure 4). This chapter gives an overview of the different steps: The computation starts with the extraction of word counts (4.2.1) and co-occurrence windows (4.2.2) from a given corpus. The information gathered is then used to compute a *centroid vector* for each LU (4.2.3) and each frame (4.2.4). In the next step, the vectors are compared with other frame vectors to compute similarities (4.2.5).

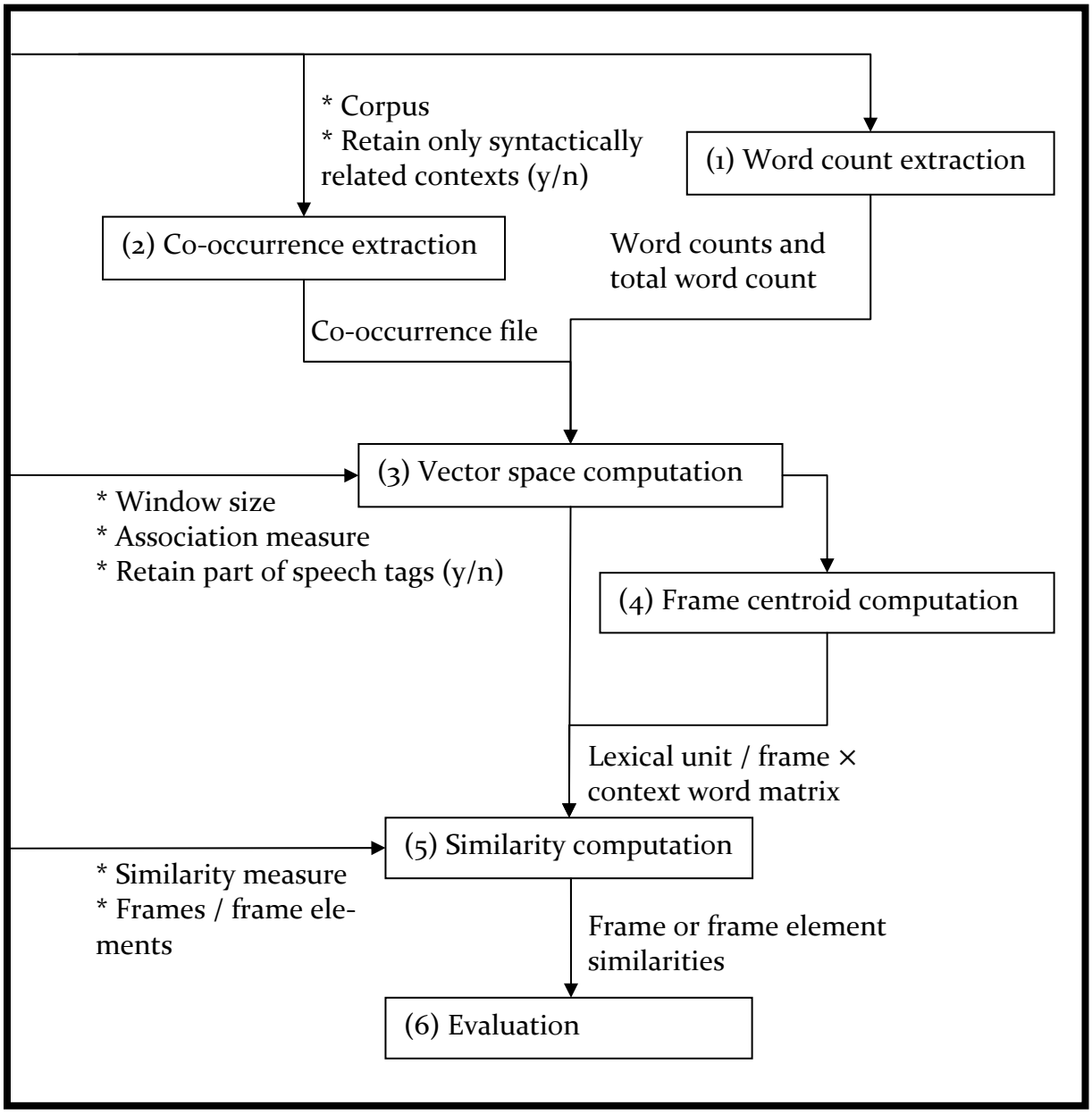

Figure 4: Overview of the system architecture

4.2.1 Extraction of word counts

In the first step, counts for all occurring words $w \in W$ in the corpus are created by processing all available text files once and setting up counters ($freq(w) = 1$) for every previously unseen word. If a previously word re-occurs, the system increases the respective word counter ($freq(w) = freq(w) + 1$). In addition to single word counts, counts for multi-word LUs $l \in L$ ($L \subset W^+$) are also considered. For this purpose, the

system extracts the respective set of LUs L from FrameNet's database beforehand. The counts extracted in this step are used later on to calculate the vectors of lexical units and their influence on each frame vector.

4.2.2 Extraction of co-occurrence window

Given the set L of all lexical units in FrameNet, the goal of this step is to extract all co-occurrence windows for each lexical unit $l \in L$. Our system starts by parsing the FrameNet database in XML format to find the set of all lexical units. The LUs are used to extract all of their co-occurring contexts C (*co-occurrence windows*) in a given corpus by processing the available text word by word.

To extract contexts for the *bag-of-words model*, a buffer stores the context of the current word in each position (by default, the 20 previous words). When encountering a LU $l_i \in L$, the current buffer is saved to an output file corresponding to the LU, i.e. the system writes out the previous context words $c_{-20\,\ldots\,-1} \in C_i$. If the last word in the context buffer is a LU, the current buffer plus the next word (i.e. the following context $c_{1\,\ldots\,20} \in C_i$ of the LU l_i) is written to the same output file. This way, our system can handle the previous and following words within the same data structure. This strategy guarantees for this step a runtime linear to the corpus size ($= O(n)$).

Before adding a word to the context buffer, we remove punctuations and normalize words and their respective part-of-speech tags. Words are converted to lower case to avoid mismatches at the beginning of a sentence, and punctuations within words are removed due to incorrect tokenization[16] (e.g. "also," → "also"). The part-of-speech tags given by RASP are mapped to the more generic tags used in FrameNet (e.g. "VBZ" → "V") to identify lexical units (cf. 4.1).

[16] Since our system expects input parsed by RASP, words in the corpus have already been tokenised and lemmatised.

The output format of this step is a list of co-occurring words along with their respective distances to the LU in whose context it was found. This format can be used in the next step to create co-occurrence matrices with different parameters; i.e. the window size and stop words can be varied to consider only selected co-occurrences or smaller windows. Figure 5 shows an example of input and the resulting format for the LU "straightforward".

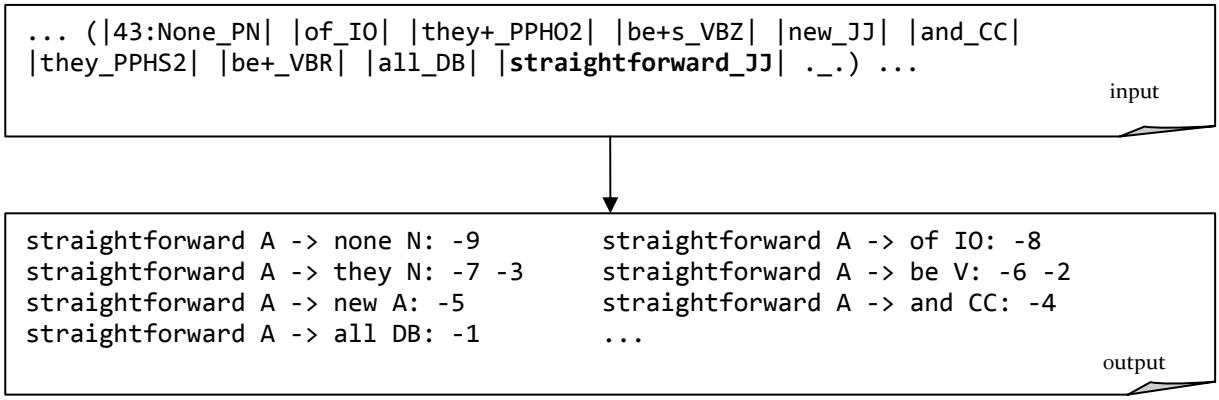

Figure 5: Input and output file format (*bag-of-words model*)

For the construction of a *dependency model*, the system writes out only those context words $c \in C_i$ that are in the same sentence and in (any) syntactic relation to a LU l_i. Instead of a string buffer, a simple graph is used here to determine the relevant context. The graph represents words as nodes, grammatical functions as (undirected) weighted[17] edges between the nodes, and is automatically constructed from the relations given in the RASP output format (cf. Figure 6).

[17] We weight all grammatical relations equally except for conjunctions which are ignored in our model. Thus "Rob ate and Bob ate" and "Rob and Bob ate" will result in the same representation in our dependency model: (Rob, eat) = 1 and (Bob, eat) = 1.

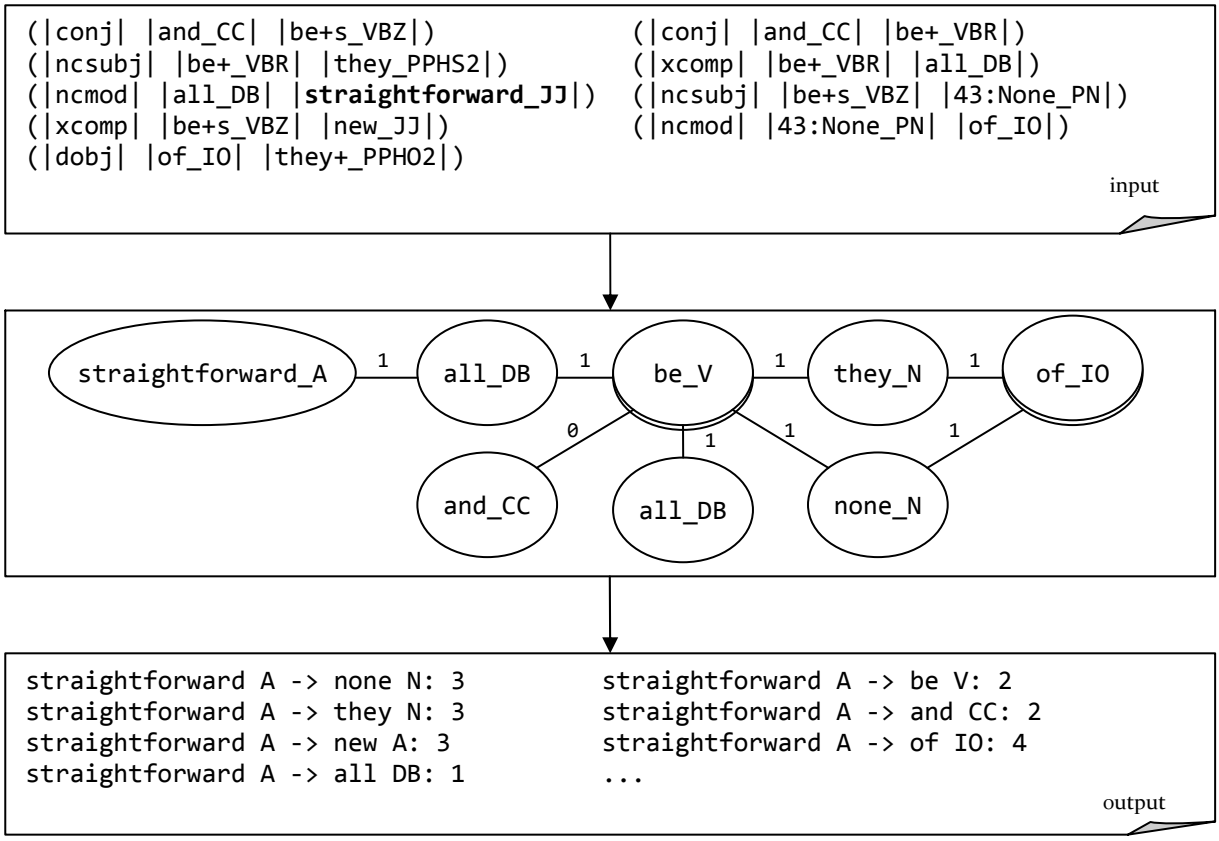

Figure 6: Input and output file format (*dependency model*)

4.2.3 Vector space computation for LUs

In this step, the system calculates the vector representation $\vec{l}$ for each lexical unit $l \in L$. The calculation uses the context windows C from the previous step (4.2.2) and the word counts $freq(w)$ from the first step (4.2.1) in order to compute vectors in the semantic space. Each dimension d in this space is labelled with a context word $c \in C$.

The selection of dimensions $D \subseteq C$ is based on a threshold. Every word below this threshold is ignored. The reason for this is that low frequency words lead to a sparse data problem, which would make the space noisier than reliable (cf. (McDonald, 2000)). Since the first vector space model by Salton et al. (1975), additional stop word lists are used to remove high frequency and function words that do not provide any

significant meaning[18]. The selection of specific dimensions is not only useful for removing noise, but also for simplifying the calculation of vector similarity since every unfiltered word adds a new dimension to the vector space. The runtime complexity of this step is linear to the number of LUs times the number of dimensions ($= O(|L| * |D|)$).

The value of a dimension $d \in D$ for the lexical unit $l \in L$ equals one of the following association measures[19]: co-occurrence frequency, conditional probability or point-wise mutual information. The actual values of these measures are computed using maximum-likelihood estimates (MLE).

Co-occurrence frequency is a simple measure that counts how often two words co-occur and is generally used as a baseline to evaluate other measures (Evert & Krenn, 2001):

$$Cooc(l,d) = p(l,d) \approx^{MLE} \frac{freq(l,d)}{corpus\_size} = \frac{freq(l,d)}{\sum_{c \in C} freq(c)}$$

Conditional probability allows for computing the probability of a lexical unit l given a specific context word d, i.e. it estimates with which probability a occurrence of d indicates a occurrence of l:

$$CP(l,d) = \frac{p(l,d)}{p(d)} \approx^{MLE} \frac{freq(l,d)}{freq(d)}$$

(Point-wise) Mutual Information accounts for the fact that words can have different frequencies and measures how significant the number of real co-occurrences is com-

[18] A list of stop words and part-of-speech tags can be found in Appendix A: Stop Word Lists.

[19] If $freq(l) = 0$ or $freq(w) = 0$, we assign the value 0 instead of the actual association measure to avoid division by zero.

pared to an estimated co-occurrence count based on their frequencies (Church & Hanks, 1989)[20]:

$$MI(l,d) = \frac{p(l,d)}{p(l) * p(d)} \approx^{MLE} \frac{freq(l,d) * corpus\_size}{freq(l) * freq(d)} = \frac{freq(l,d) * \sum_{c \in C} freq(c)}{freq(l) * freq(d)}$$

In order to find the best suitable model for frame semantics, the association measure and additional options can be parameterized. Further parameters include the window dimension specifying how many context words before and after the LU are taken into account, a file with part-of-speech tags and words to skip, and a lower bound for word frequencies.

The output of this step is a file with a dense matrix containing vectors for all LUs and two files with labels (cf. Figure 7). The first line of the main file gives information about the matrix dimensions, i.e. number of rows $|L|$ and number of columns $|D|$. All further lines represent the values of one vector each. The two label files contain labels for both rows (LUs L) and columns (dimensions D) of the calculated matrix.

```
8310 10
0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 17.0 0.0 0.0 0.0 0.1 0.0
0.0 0.0 7.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 2.5 0.0 0.0 0.0 0.0 0.0
0.0 5.3 0.0 0.0 0.0 0.0 0.0 0.0 0.6 0.0
0.0 0.0 0.0 6.2 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 2.8 0.0 0.0 0.0 0.0
0.0 9.5 0.0 0.0 0.0 0.0 0.0 12.0 0.0 0.0
0.0 0.0 0.0 0.0 6.8 0.0 0.0 0.0 0.0 0.0
...
```

Figure 7: Matrix file format

[20] Note that we left out the logarithm in the actual formula to avoid negative values as some of our similarity measures (cf. 4.2.5) require positive input values.

4.2.4 Vector space computation for frames

The computation of centroid frame vectors $\vec{f}$ is based on the lexical units vectors $\vec{l}$. For each frame $f \in F$, we compute its representation based on its lexical units $L(f)$. In particular, the dimension i of a frame centroid is calculated by summing up over the respective values of its associated lexical units $l \in L(f)$:

$$f_i = \sum_{l \in L(f)} l_i * \underbrace{\frac{freq(l)}{freq(f)}}_{w}$$

The weighting factor w for each value is defined as the relative number of occurrences that one LU has compared to all LUs from the considered frame, so that the resulting frame vector represents an average distribution over all co-occurrences of the associated LUs. The weighting ensures that low frequency LUs have a smaller impact on the frame centroid than high frequency LUs since these typically have a more robust vector representation. The runtime complexity of this step is (approximately) linear to the number of frames times the number of dimensions ($\approx O(|F| * |D|)$).

The output format of this step is equal to the output of the previous step. The only difference is that the output file contains vectors for frames instead of LUs, thus the matrix has a smaller amount of rows.

4.2.5 Similarity computation

In the similarity computation step, the previously created matrices are used to compare vector representations to each other. This step can be used to calculate the similarity between the representations of two LUs, between a LU and a frame, and between frames. For every pair of vectors $(\vec{x}, \vec{y})$, our system can apply different measures that have been proven to be useful for calculating semantic similarity (Mohammad & Hirst, 2006). A measure which is often used (e.g. by Schütze and Pederson (1997), and Yo-

shida et al. (2003)) is cosine similarity, which calculates the cosine of the angle between two vectors.

- Cosine similarity

$$\text{sim}_{\cos}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| * |\vec{y}|}$$

The main advantage of this measure is that it takes both real values and total vector lengths into account. This is useful because the vector lengths can contribute as a normalization factor when a matrix does not contain already normalized values. Normalization is an important factor in ensuring that similarity computations are not biased by word frequency. This is especially true for the comparison of two LUs that have a very different frequency (for example, the infrequent verb "sop" and the frequent verb "soak"). However, if the given matrix contains values resulting from an association measure that already considers word frequency, further normalization can be skipped. Some of the following similarity measures, which we additionally implemented, do not perform any normalization:

- Jaccard coefficient (Jaccard, 1908)

$$\text{sim}_{\text{jaccard}}(\text{x}, \text{y}) = \frac{|x \cap y|}{|\text{x} \cup \text{y}|}$$

The original Jaccard coefficient computes the similarity of two sets by comparing their common elements to the number of elements in the union of both sets. While elements in this measure are treated as binary values, it is also possible to use their actual values in a modified version as proposed by Dagan et al. (1993):

- Modified Jaccard coefficient

$$\text{sim}_{\text{jaccard}}(\vec{x}, \vec{y}) = \sum_i \frac{\min(x_i, y_i)}{\max(x_i, y_i)}$$

Another measure we want to test is Jensen-Shannon divergence, a symmetric similarity measure based on the Kullback-Leibler divergence:

- Jensen-Shannon divergence (Lin J. , 1991)

$$sim_{JSD}(\vec{x},\vec{y}) = \sum_{\{i\,|x_i>0,y_i>0\}} y_i * \log\frac{y_i}{½ * (x_i + y_i)}$$

Finally, the following two similarity measures are derived from information theory and are used to compare point-wise mutual information[21]:

- Lin's similarity measure (Lin D. , 1998)

$$sim_{Lin}(\vec{x},\vec{y}) = \frac{\sum_{\{i\,|x_i>0,y_i>0\}}(x_i + y_i)}{\sum_{\{i\,|x_i>0,\}} x_i + \sum_{\{i\,|y_i>0\}} y_i}$$

- Saif's similarity measure (Mohammad & Hirst, 2006)

$$sim_{saif}(\vec{x},\vec{y}) = \frac{2 * \sum_{\{i\,|x_i>0,y_i>0\}} min(x_i, y_i)}{\sum_{\{i\,|x_i>0,\}} x_i + \sum_{\{i\,|y_i>0\}} y_i}$$

[21] Note that we removed the relation variable from both Lin and Saif's similarity measures since relations are not explicitly stored in our matrices.

Chapter 5

Experiments

The experiments described in this chapter were performed with different semantic space models to evaluate the impact of various parameters. All models were built with the British National Corpus (BNC; Burnard, 2000), a 100 million word text corpus of written and spoken English, and have the following properties:

- **Model type**: Bag-of-words model or dependency model (cf. 3.3)
- **Association measure**: co-occurrence frequency, conditional probability or mutual information (cf. 4.2.3)
- **Distance weighting:** no weighting or $\frac{1}{d}$-weighting[22]
- **Distance measure:** cosine or modified Jaccard coefficient[23] (cf. 4.2.5)
- **Maximum distance:** 5, 10, or 20 for context windows; 1 or 3 for syntactic relations[24]

The following tests are conducted with a fixed model-size of about 4,000 dimensions representing all words occurring at least 2,000 times in the BNC (excluding stop

[22] When using $\frac{1}{d}$-weighting, words in position d of a context window are counted as $1/|d|$ co-occurrences with the lexical unit.

[23] We use only these two measures because other measures performed either equally well or worse.

[24] For the syntax-based models, a maximum distance of 1 means that we only consider words directly related to the LU; 3 means that all words are considered that are related to a LU given up to two other words inbetween.

words[25]). As pointed out by McDonald (2000), 100-2,000 dimensions are typically used in a vector space model unless some kind of dimensionality reduction is applied. However, we chose a higher number of dimensions as this proved to be more robust, e.g. Schütze and Pedersen (1997) use 3,000 dimensions, and Sahlgren and Cöster (2004) report best performance with 5,000 dimensions.

5.1. Similarity Experiment

The purpose of this first experiment is to confirm our hypothesis that LUs of the same frame have a similar distributional representation. We expect to obtain higher similarity values between two LUs of the same frame than between LUs randomly chosen from different frames. By analyzing the results of different configurations, we also gain insight into which vector spaces are better suited for frame semantics than others.

To do so, we create a true set (TS) that contains all pairs of LUs that evoke a mutual frame and a control set (CS) of pairs of LUs not in the same frame. The TS is defined as follows:

$$TS = \{< l_1, l_2 > \in L^2 \mid \exists f \in F. \{l_1, l_2\} \subseteq L(f)\}$$

The CS is formed by as many randomly chosen LU pairs that do belong to a mutual frame:

$$CS = \{< l_1, l_2 > \in L^2 \mid \nexists f \in F. \{l_1, l_2\} \subseteq L(f)\}\ s.t. |TS| = |CS|$$

For each LU pair in the true and control set (around 350.000 in total), we compute the similarity using the measures described in 4.2.5. In order to evaluate the actual performance of these similarities, we create a *ROC curve* (Fawcett, 2006) that shows how well a statistical measure is able to distinguish between the true set and the control set.

[25] The stop word list that has been used to filter out noisy words and part-of-speech tags can be found in Appendix A: Stop Word List.

This is done by modelling a curve with the rate of true positives (TP) on one axis and the rate of false positives (FP) on the other axis for each similarity value (cf. Figure 8).

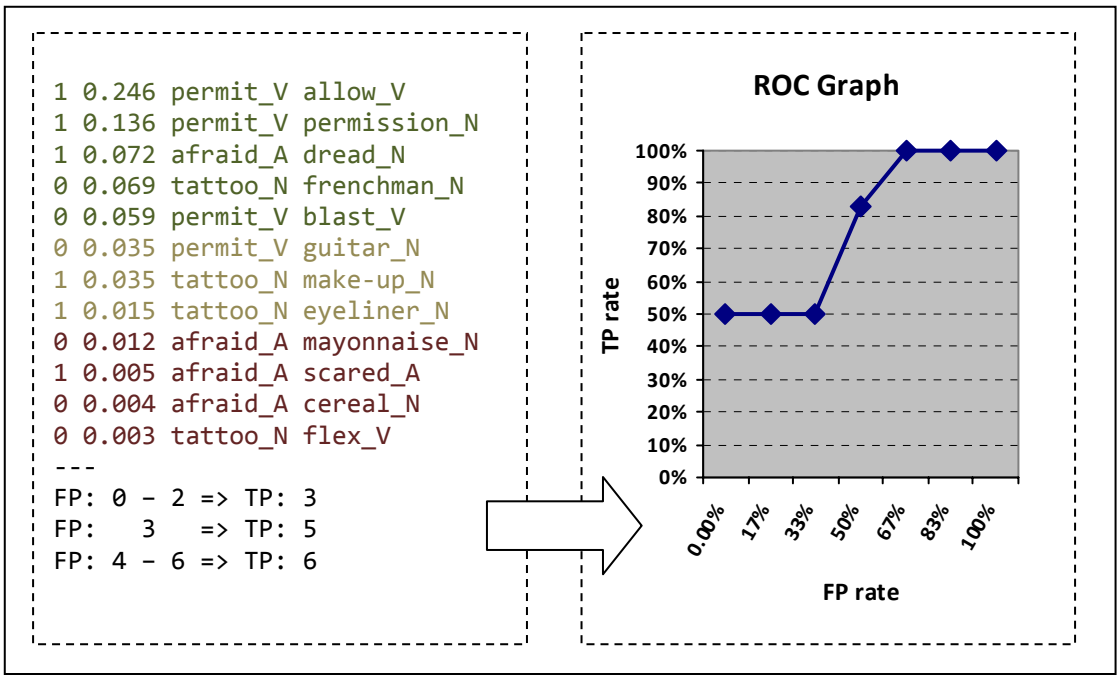

Figure 8: Similarities and ROC graph representation

We also compute the classification accuracy at each point in the ROC curve in order to find the similarity threshold for the best classification performance (*best accuracy point*). A measurement often used for comparing ROC curves of different classifiers is the (total) Area under the ROC curve (AROC). This is a simple means of comparison based on the fact that better performing methods will have a smaller FP rate (higher TP rate), meaning that the curve is more to the left (up) and has a higher AROC (cf. Figure 9). The minimum AROC is always 0% (the TP rate is 0% at any FP rate), while the maximum is 100% meaning that every element is classified correctly.

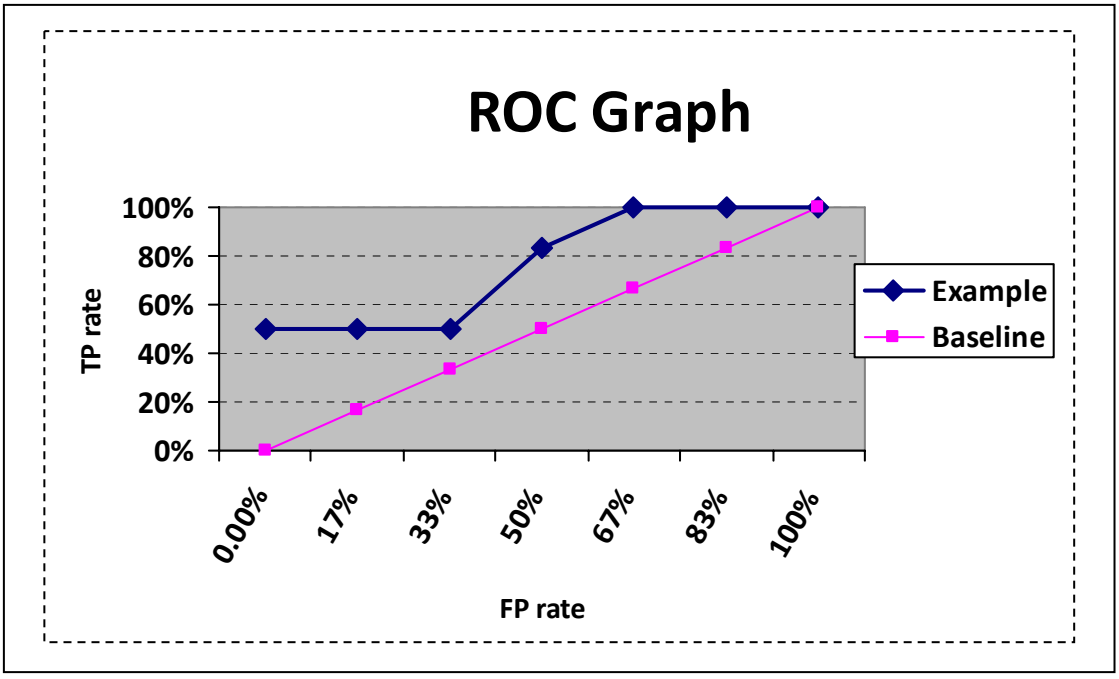

Figure 9: Comparison between the baseline and a sample of our results

5.1.1 Results

We compute the performance of all models by means of best accuracy and AROC. Since this is a binary classification task, the random baseline has both an accuracy and AROC of 50%. Tables 3-8 below give an overview of the resulting values in percentage for all models tested (best values compared to the baseline in parenthesis).

dependency	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	CP	CP	CP	MI	MI	MI
Window size	1	3	3	1	3	3	1	3	3
Weighting	O	O	X	O	O	X	O	O	X
AROC	58.6	53.9	55.8	58.6	54.0	55.9	62.5	63.0	**63.1** (+13.1)
Best Accuracy	57.5	53.7	55.3	57.0	53.8	55.3	61.3	61.4	**61.8** (+11.8)

Table 3: LU-LU results for dependency models (cosine measure)

dependency	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	CP	CP	CP	MI	MI	MI
Window size	1	3	3	1	3	3	1	3	3
Weighting	O	O	X	O	O	X	O	O	X
AROC	59.9	57.0	57.5	**64.0** (+14.0)	58.4	62.0	63.5	59.7	62.8
Best Accuracy	57.3	54.6	54.7	**61.4** (+11.4)	56.7	60.0	60.7	57.9	60.5

Table 4: LU-LU results for dependency models (Jaccard measure)

bag-of-words	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	Cooc	Cooc	Cooc	CP	CP	CP
Window size	5	5	10	10	20	20	5	5	10
Weighting	O	X	O	X	O	X	O	X	O
AROC	50.7	49.8	51.8	50.4	53.9	51.0	50.7	49.7	51.8
Best Accuracy	51.7	51.3	52.5	51.7	53.8	52.2	51.7	51.3	52.5

Table 5: LU-LU results for bag-of-words models 1-9 (cosine measure)

bag-of-words	10	11	12	13	14	15	16	17	18
Association measure	CP	CP	CP	MI	MI	MI	MI	MI	MI
Window size	10	20	20	5	5	10	10	20	20
Weighting	X	O	X	O	X	O	X	O	X
AROC	50.4	53.8	51.7	54.2	52.0	56.2	52.5	**60.3** (+10.3)	53.5
Best Accuracy	51.8	53.7	52.7	53.9	53.2	55.1	53.6	**58.1** (+8.1)	54.4

Table 6: LU-LU results for bag-of-words models 10-18 (cosine measure)

bag-of-words	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	Cooc	Cooc	Cooc	CP	CP	CP
Window size	5	5	10	10	20	20	5	5	10
Weighting	O	X	O	X	O	X	O	X	O
AROC	53.7	53.1	54.2	53.4	54.5	53.5	52.3	52.6	53.2
Best Accuracy	53.1	52.8	53.4	52.9	53.5	53.0	52.1	52.5	52.8

Table 7: LU-LU results for bag-of-words models 1-9 (Jaccard measure)

bag-of-words	10	11	12	13	14	15	16	17	18
Association measure	CP	CP	CP	MI	MI	MI	MI	MI	MI
Window size	10	20	20	5	5	10	10	20	20
Weighting	X	O	X	O	X	O	X	O	X
AROC	53.6	**55.0** (+5.5)	54.3	52.2	51.7	52.7	52.0	53.6	52.4
Best Accuracy	53.1	**54.3** (+4.3)	54.3	51.7	51.8	52.1	52.1	52.8	52.4

Table 8: LU-LU results for bag-of-words models 10-18 (Jaccard measure)

5.1.2 Results Analysis

The results of this experiment show that most of our selected models yield a higher AROC and accuracy than the baseline. In fact, only one configuration (*dependency model 8*) yields worse performance with respect to the AROC. Figure 10 shows the ROC curves for the two best performing configurations and the baseline.

When comparing the curves, it is clear that our models outperform the baseline for most of the cases. However, there seem to be a number of LUs from the test set that are not more similar to each other than the pairs from the control set. Two possible reasons behind this are ambiguity (LUs evoking multiple frames probably have a different representation than non-ambiguous words) and data sparseness (the representation of infrequent words is more sparse than that of frequent words).

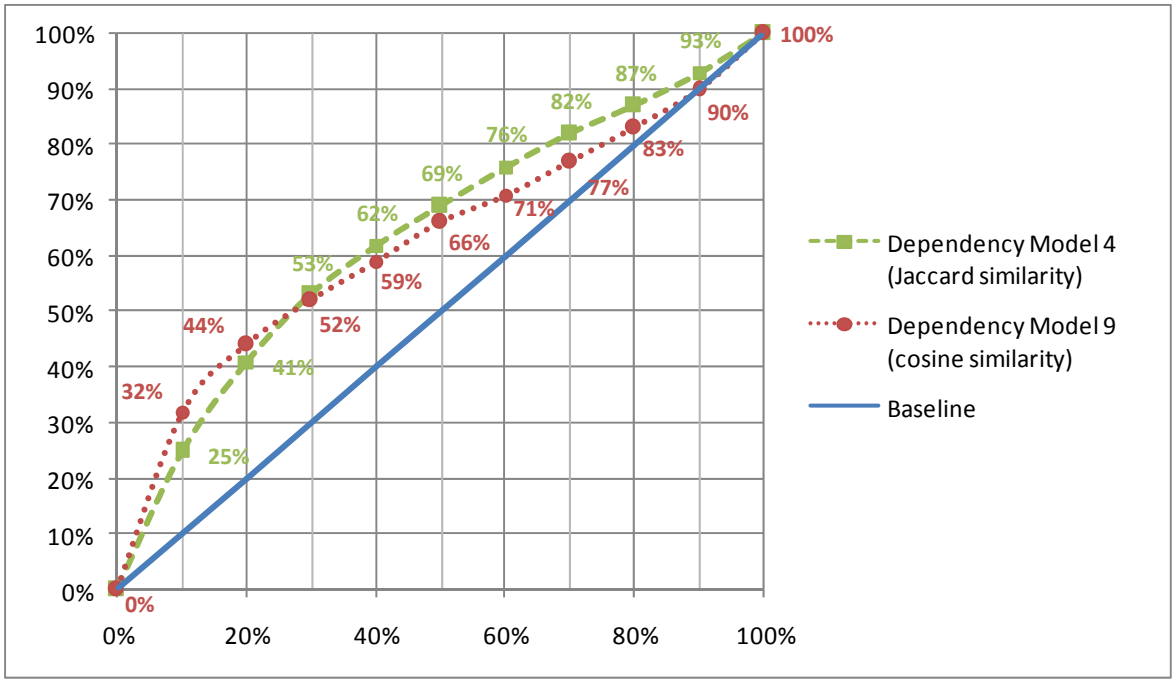

Figure 10: The best performing models and random baseline (ROC curves)

We discuss the parameter's influence on best accuracy and AROC in the following sections:

Model type: Both syntax and context-based matrices performed equally well, though the best results are achieved with the *dependency model 4*.

Association Measure and **Distance Measure:** Overall, matrices that are computed using co-occurrence frequency as an association measure showed the worst results. We believe that the main reason for this is that the frequencies of the LUs are distributed over a decent range. Consequently, more frequent LUs have a higher average association measure than low frequency LUs and most frames have LUs from different frequency ranges leading to low similarity results. The other two association measures yield better similarity results: Conditional probability appears to be the best association measure for computing jaccard similarity, whereas Mutual Information performs best for cosine similarity. The best results for dependency models are achieved with jaccard (*model 4*), whereas it is cosine for bag-of-words models (*model 17*).

Window Size and **Distance Weighting:** Weighting the distance between two words improves the results of dependency-based matrices, though it does not do so for context-based matrices. Apparently, words that are directly related to a target syntactically comprise more frame semantic information than other words in context. This can also be seen by the fact that a maximum distance of 1 yields better results for a dependency model than a distance of 3. In a bag-of-words model, however, syntactically related words can be found within any distance to the target word, e.g.

"[We $_{\text{SUBJ/Seller}}$] do the maintenance instead of just **selling** [the plants $_{\text{OBJ/Goods}}$]."

Thus all words in a context window potentially have the same importance. Since syntactically related words are at the same time often frame element fillers, these observations confirm our intuition that frame elements are important indicators for a frame.

5.2. Leave-One-Out Experiment

This second experiment aims to test the possibility of assigning an "unknown" LU to the right frame by comparing vector representations in the semantic space. Our intuition for this experiment is that the similarity between a LU and the evoked frame should be higher than the similarity between a LU and any other frame. We expect the results of this evaluation to be a useful indicator for the possibility of automatically expanding FrameNet. One way in which these results can be used is to suggest the most similar frames for unknown LUs to a human annotator in order to reduce the time needed when assigning LUs to frames.

The experiment consists of comparing each LU to all frame centroid vectors to retrieve the most probable frame(s): $f_{best}(l) = \arg\max_{f \in F} sim(\vec{l}, \vec{f})$

In order to avoid biased results, our system re-computes the frame centroid in each comparison without the considered LU.

5.2.1 Results

We measure the performance of a model by counting how many LUs are assigned to the frame they actually evoke (*precision*). To gain better insight into the differences between LUs and their frame representations, we compute the rank r_l of the correct frame f_l for each LU l and the median rank $\tilde{r}$ over all LUs (i.e. the middle value of an ordered set R that contains the ranks for all LUs):

$$r_l = |\{f \in F \mid l \notin L(f) \wedge sim(\vec{l}, \vec{f}) > sim(\vec{l}, \vec{f_l})\}|$$

$$R = < r_1, r_2, \ldots, r_n > \; s.t. \; r_i \geq r_{i+1}$$

$$\tilde{r} = \begin{cases} r_{|L|/2} & |L| \; odd \\ 1/2 \left(r_{|L|/2} + r_{(|L|+1)/2}\right) & |L| \; even \end{cases}$$

During the experiment, we compute the similarity between all 8.310 LUs and the 794 frames currently in the FrameNet database. The simplest baseline for this approach is a random frame assignment for each LU, resulting in a precision of less than 1% and a median rank of 397 (half the total number of frames). A more informed baseline that assigns all lexical units to the most probable frame from a naive point of view, i.e. the frame with the most lexical units, results in 179 right classifications out of 8.310 (about 2% precision). Tables 9-14 give an overview of the results with our models (best values compared to baseline in parenthesis).

dependency	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	CP	CP	CP	MI	MI	MI
Window size	1	3	3	1	3	3	1	3	3
Weighting	O	O	X	O	O	X	O	O	X
Precision	4%	11%	9%	9%	13%	11%	18%	**25%** (+23%)	**20%**
Median Rank	121	105	116	81	68	76	16	**8** (+389)	12

Table 9: LU-frame results for dependency models (cosine measure)

dependency	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	CP	CP	CP	MI	MI	MI
Window size	1	3	3	1	3	3	1	3	3
Weighting	O	O	X	O	O	X	O	O	X
Precision	7%	7%	7%	16%	20%	18%	21%	23%	**23%** (+22%)
Median Rank	218	267	254	42	21	28	17	14	**12** (+385)

Table 10: LU-frame results for dependency models (Jaccard measure)

bag-of-words	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	Cooc	Cooc	Cooc	CP	CP	CP
Window size	5	5	10	10	20	20	5	5	10
Weighting	O	X	O	X	O	X	O	X	O
Precision	9%	8%	9%	8%	12%	9%	9%	8%	10%
Median Rank	139	159	128	156	98	144	104	126	93

Table 11: LU-frame results for bag-of-words models 1-9 (cosine measure)

bag-of-words	10	11	12	13	14	15	16	17	18
Association measure	CP	CP	CP	MI	MI	MI	MI	MI	MI
Window size	10	20	20	5	5	10	10	20	20
Weighting	X	O	X	O	X	O	X	O	X
Precision	8%	14%	9%	17%	15%	19%	15%	**23%** (+21%)	17%
Median Rank	118	63	106	32	50	25	45	**13** (+384)	38

Table 12: LU-frame results for bag-of-words models 10-18 (cosine measure)

bag-of-words	1	2	3	4	5	6	7	8	9
Association measure	Cooc	Cooc	Cooc	Cooc	Cooc	Cooc	CP	CP	CP
Window size	5	5	10	10	20	20	5	5	10
Weighting	O	X	O	X	O	X	O	X	O
Precision	5%	5%	4%	4%	5%	4%	8%	7%	9%
Median Rank	288	284	287	283	290	285	82	100	69

Table 13: LU-frame results for bag-of-words models 1-9 (Jaccard measure)

bag-of-words	10	11	12	13	14	15	16	17	18
Association measure	CP	CP	CP	MI	MI	MI	MI	MI	MI
Window size	10	20	20	5	5	10	10	20	20
Weighting	X	O	X	O	X	O	X	O	X
Precision	7%	14%	7%	9%	10%	10%	11%	**15%** (+13%)	**12%**
Median Rank	104	41	100	68	70	58	62	**34** (+363)	53

Table 14: LU-frame results for bag-of-words models 10-18 (Jaccard measure)

5.2.2 Results Analysis

The results of this experiment show that all models are better suited to finding the correct frame of a lexical unit than the random baseline. Though our system only classifies 25% of the lexical units correctly, the median rank of the right solution reveals that a correct frame can be found within the 8 most similar results in more than 50% of the cases. When extending the model to 30.000 dimensions, the top precision improves to 27% correct frame assignments. This outcome indicates the possibility of considerably assisting frame assignment for new lexical units. Figure 11 shows the precision of the best performing model (and an extended model with 30.000 dimensions) when also considering the top-20 most similar frames for each lexical unit.

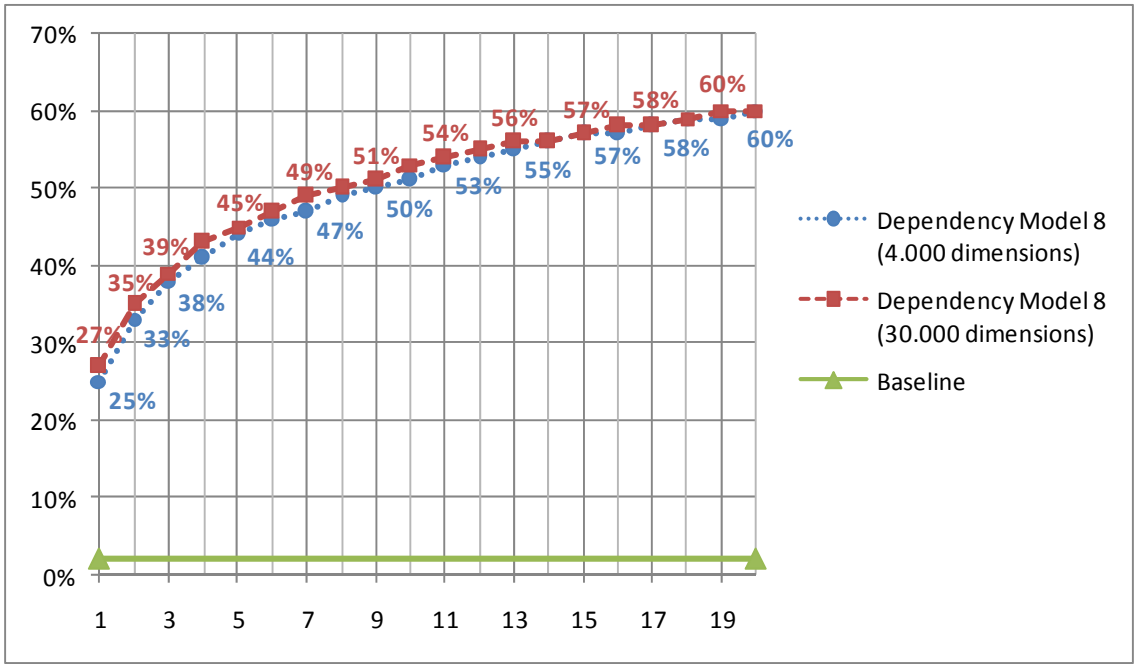

Figure 11: Top-20 precision of the best performing model

Compared to the baseline, the results obtained from this experiment are better than those from the first experiment. The reason for this is that frame centroids are more reliable than LU vectors because they do not suffer from problems like ambiguity and data sparseness (typically a frame consists of multiple LUs not all of which are polysemous or infrequent).

A closer look into our results shows that frames which were ranked higher are many times related to the correct frame. Indeed, in cases where the correct frame is among the top ten results, higher ranked results quite often are plausible substitutes for those actually assigned in the FrameNet database. The following list gives a few examples indicating the correct frame according to FrameNet and those assigned by our system[26]:

[26] An overview of 100 assignments made with our best performing model can be found in Appendix B: Most Similar Frames.

- carve V (CREATE_REPRESENTATION) → 1. SHARPNESS, 2. CUTTING, …

Even though the verb "carve" potentially describes an event in which a "**Creator** produces a physical object which is to serve as a **Representation** of an actual or imagined entity or event", in our opinion, the actual process intuitively fits better with the frame CUTTING. Between FrameNet version 1.3 and the current state, which can be viewed online, "carve" actually got assigned to the CUTTING frame as well. The reason why SHARPNESS is closely related to this word in our model can be seen by the fact that "sharpness" is a required property for all processes described with the CUTTING frame.

- guerrilla N (PEOPLE_BY_VOCATION) → 1. MILITARY, 2. TERRORISM, …

"Guerrilla" is another good example of a word that fits the definition of multiple frames. In the current version of FrameNet, however, it evokes only the frame PEOPLE_BY_VOCATION though, intuitively, the frames proposed by our model are equally valid: MILITARITY describes "some **Possessor**, either a nation, institution, or private individual, [who] controls a Force (…)" and TERRORISM is defined as an event in which a "**Terrorist** commits a violent or otherwise harmful **Act** upon a **Victim** in order to coerce or terrorize a government or populace."

- caravan N (BUILDINGS) → 1. VEHICLE, 2. BUILDINGS, …

According to our best performing model, the frame VEHICLE is most similar and BUILDINGS is the second most similar frame to the lexical unit caravan. Conceptually, this makes sense since a caravan describes an object that "form[s] an enclosure and provide[s] protection" (BUILDINGS) and that at the same time can be used for "the purpose of transportation" (VEHICLE). However, as FrameNet only contains BUILDINGS as a frame for caravan, we believe that this is a good example showing the expansion potential that could be achieved through this approach.

5.3. Discussion

The results of both our experiments show that semantic spaces are, to some extent, a suitable means of representing frame semantic meaning and potentially useful in assisting the expansion of FrameNet. Moreover, the results obtained from our leave-one-out evaluation are useful indicators of additional frames evoked by existing lexical units in the FrameNet database.

Although our results rank below state-of-the-art methods (Erk, 2005; Burchardt, Erk, & Frank, 2006) in terms of precision, our approach nevertheless clearly outperforms other systems through coverage in that it does not depend on labelled training data or additional semantic resources. For example, Detour (Burchardt, Erk, & Frank, 2005) heavily depends on the words contained in WordNet, thus only covering 87% of the cases appearing in the annotated FrameNet corpus. Moreover, this annotated corpus itself only provides training data for around 60% of all lexical units currently in the FrameNet lexicon. In contrast to this, a semantic space representation as proposed in our work can be computed for any arbitrary word (labelled or unlabelled) that occurs in the corpus. Moreover, this approach is also applicable for other languages, for which a small set of lexical units are already assigned to frames.

To confirm these observations outside the range of FrameNet, we conducted the same experiments in the SALSA framework (Erk, Kowalski, Padó, & Pinkal, 2003), a German frame-based lexical semantics resource, as well as the German part of the Europarl corpus (Koehn, 2005). With a top precision of 26%, the best results[27] of this small-scale experiment are comparable to the findings described in the previous section. When considering the top-20 most similar frames, 53% of correct frame assignments can be made (cf. Figure 12).

[27] Note that we only evaluated the bag-of-words model because RASP cannot be used for German input.

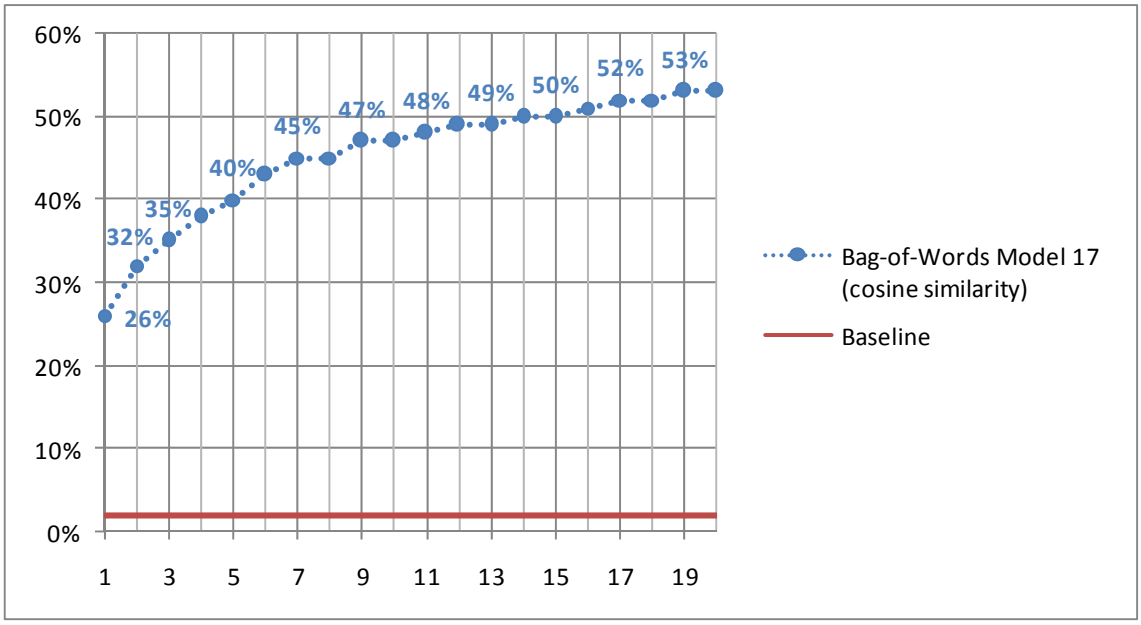

Figure 12: Top-20 precision of the best performing model (SALSA)

Chapter 6

Conclusions

This work attempts to model frame semantic meaning in a semantic space. In contrast to previously computed models, e.g. (Burchardt, Erk, & Frank, 2005) and (Fung & Chen, 2004), the goal of our study was to use a minimum of available resources in order to achieve the highest possible coverage. Though our results do not yet fall in line with precision achieved by state-of-the-art systems (e.g. (Erk, 2005)), our approach compensates in its outstanding coverage.

Our results indicate that a syntax-driven approach is better suited for the actual construction of a semantic space model. Presumably, this fact is based on the observation that roles in a frame are typically filled by syntactically related words. Any other words which do not provide relevant information for the frame are filtered out in a purely syntax-based model. Based on the overall results of our experiments, our methods show promise for extending the existing yet incomplete FrameNet database.

In the following sections, we discuss the limitations of this approach, potential improvements and applications of the results.

6.1. Known Issues

One of the main issues that we encountered during the evaluation is that very infrequent LUs such as "to sop" (13 occurrences in 100 million words) have a very sparse vector representation. This is problematic because only few dimensions are available for comparison, and if the values in those dimensions are not similar to those in the correct frame vector, the assignment fails. To verify this hypothesis, we ran a number of experiments with thresholds on the LU frequency, confirming our intuition that data sparseness is indeed a problem in our approach (cf. Figure 13). One possibility to make this process more robust is to use smoothing techniques in order to fill the other dimensions with some sort of backing-off values.

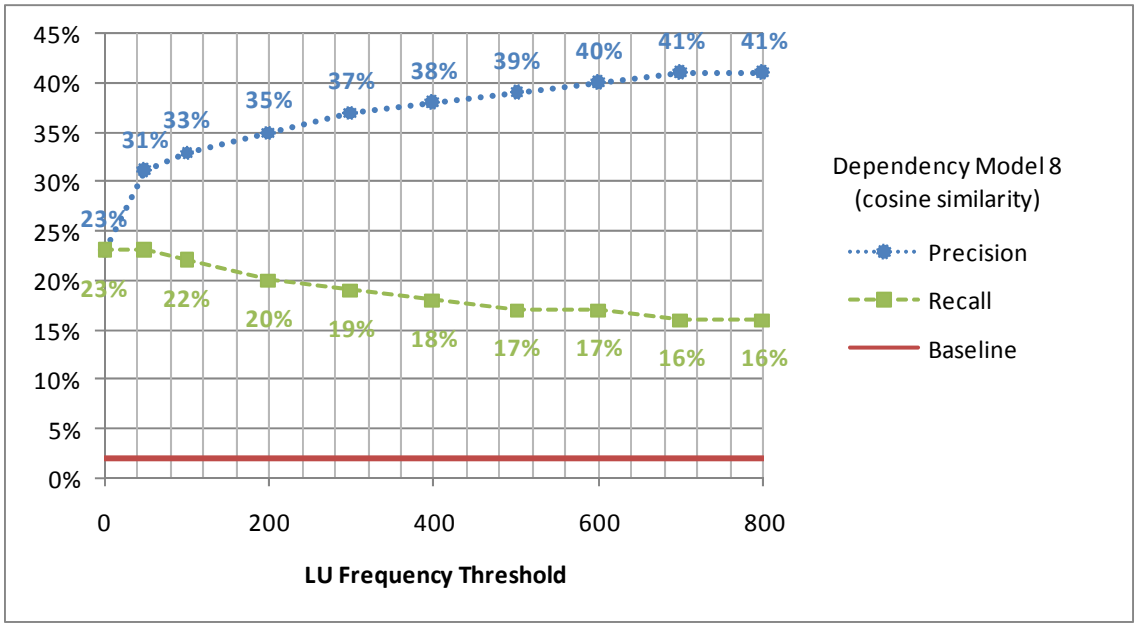

Figure 13: Top-1 precision and recall with thresholds on LU frequencies

Another issue is based on the fact that our syntax-based model currently treats all syntactic relations the same way. This leads to the problem that two different frames sometimes have very similar vector values in some dimensions, even though the same word never occurs in the same relation to a LU. One example for this is the pair of frames FOOD and INGESTION which are often evoked in contexts containing words that

refer to eatable objects. However, in the case of FOOD, these words typically modify a lexical unit (e.g. [cheese $_{MOD}$] **sandwich**), while in the case of INGESTION, they occur as objects of a LU (e.g. **eat** [cheese $_{OBJ}$]). In order to diversify the representations of both frames, one could extend the syntax-based model to have different dimensions for various relations (cf. for example (Padó, 2007)). Another possibility is to use multiple representations for different part-of-speech tags, so that "cheese" would only have a noun representation, while "eat" would have a verb representation. Following that, each frame could have a centroid vector for each part-of-speech tag to achieve a more robust performance across different lexical categories.

The incompleteness of FrameNet is another problematic aspect of our approach. Though we do not expect every word to be in the lexicon, we assumed for our experiments that the existing words in the lexicon are assigned to all frames they potentially evoke. It turned out, however, that this assumption does not hold. For example, the verb "carve" evokes but is not assigned to the frame CUTTING and the noun "caravan" is not assigned to the frame VEHICLE in the current version of FrameNet. Not only does this fact bias our evaluation results, but it influences the quality of our frame representations.

6.2. Future Work

As discussed in the previous chapter, the results of our approach are promising for further research both in the direction of pursuing further improvements of this model and its application in the actual development of FrameNet. One possibility for the former is to intelligently handle data sparseness and representation issues as discussed in the previous section. However, even without additional work, our methods can be immediately applied in guiding the frame assignment process for new lexical units.

One of the key features of this model, as stated before, is its independence from additional semantic resources. This advantage makes our methods particularly interesting for extending frame semantic resources in languages that are not as rich in resources as

English. In fact, a small set of frame-assigned lexical units and an un-annotated text corpus in the target language are sufficient to build a context-based frame model as suggested in our work. Following recent trends in cross-lingual projection (Padó, 2007), another possibility to build a frame semantic resource for another language is to take our existing model for English and transfer it to a target language using an aligned bilingual corpus. Based on this corpus, a bilingual vector space model can be computed (Pitel, 2008) containing meaning representations for words in both languages. Using the English part of those representations, the model can be directly compared to our results leading to implicit similarity results for the candidate words in the target language.

References

Baker, C. F., & Sato, H. (2003). The FrameNet Data and Software. *Poster and Demonstration at ACL.* Sapporo, Japan.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association of Computational Linguistics.* Montreal, Canada.

Berry, M. W. (1992). Large Scale Sparse Singular Value Computations. *The International Journal of Supercomputer Applications , 6* (1), 13-49.

Boas, H. C. (2002). Bilingual FrameNet Dictionaries for Machine Translation. *Proceedings of the Third International Conference on Language Resources and Evaluation.* Las Palmas, Spain.

Briscoe, E., Carroll, J., & Watson, R. (2006). The Second Release of the RASP System. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions.* Sydney, Australia.

Burchardt, A., & Frank, A. (2006). Approximating Textual Entailment with LFG and FrameNet Frames. *Proceedings of the Second Recognising Textual Entailment Workshop.* Venice, Italy.

Burchardt, A., Erk, K., & Frank, A. (2005). A WordNet Detour to FrameNet. (B. Fisseni, H.-C. Schmitz, B. Schröder, & P. Wagner, Eds.) *Computer Studies in Language and Speech , 8* (Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen), 408-421.

Burnard, L. (2000). *Reference Guide for the British National Corpus.* Retrieved from the BNC Website: http://www.hcu.ox.ac.uk/BNC/World/html/urg.html

Carroll, J., & McCarthy, D. (2000). Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities. Senseval Special Issue , 1-2* (34), 109-114.

Choi, F., Wiemer-Hastings, P., & Moore, J. (2001). Latent Semantic Analysis for text segmentation. *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing.* Seattle, Washington, USA.

Church, K. W., & Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.* Vancouver, British Columbia, Canada.

Dagan, I., Marcus, S., & Markovitch, S. (1993). Contextual Word Similarity and Estimation from Sparse Data. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.* Columbus, Ohio, USA.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science , 41 (6)*, 391-407.

Erk, K. (2005). Frame assignment as word sense disambiguation. *Proceedings of the 6th International Workshop on Computational Semantics (IWCS 6).* Tilburg, Netherlands.

Erk, K., Kowalski, A., Padó, S., & Pinkal, M. (2003). Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics.* Sapporo, Japan.

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics.* Toulouse, France.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters , 27* (8 - Special issue: ROC analysis in pattern recognition), 861-874.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* MIT Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech , 280*, 20-32.

Fleischman, M., Kwon, N., & Hovy, E. (2003). Maximum Entropy Models for FrameNet Classification. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.* Edmonton, Canada.

Fung, P., & Chen, B. (2004). BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. *Proceedings of the 20th International Conference.*

Gildea, D., & Jurasfky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistic , 28 (3)*, 496-530.

Harris, Z. S. (1968). Distributional Structure. In J. J. Katz (Ed.), *The Philosophy of Linguistics* (pp. 26-47). New York: Oxford University Press.

Hinrich Schütze, J. O. (1997). A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management , 33* (3).

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles 44* .

Karlgren, J., & Sahlgren, M. (2001). From words to understanding. In Y. Uesaka, P. Kanerva, & H. Asoh (Eds.), *Foundations of real-world intelligence* (pp. 294-308). CSLI Publications.

Klavans, J. L., & Kan, M.-Y. (1998). The Role of Verbs in Document Analysis. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association of Computational Linguistics.* Montreal, Canada.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit X.* Phuket, Thailand.

Landauer, T. K., & Dumai, S. T. (1997). A Solution to Plato's Problem: the Latent Semantic AnalysisTheory of Acquisition, Induction and Representation of Knowledge. *Psychological Review , 104 (2)*, 211-240.

Levin, B. (1993). *English Verb Classes and Alternations.* Chicago University Press.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *In Proceedings of the 15th International Conference on Machine Learning.* San Francisco, USA.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. *Proceedings of the 17th Conference on Computational Linguistics.* Montreal, Canada.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory , 37* (1), 145-151.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers , 2* (28), 203-208.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. *Proceedings of the 17th Annual Conference of the Cognitive Science Society (CogSci'95).*

McCarthy, D., Carroll, J., & Preiss, J. (2001). Disambiguating Noun and Verb Senses Using. *Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01.* Toulouse, France.

McDonald, S. (2000). Environmental Determinants of Lexical Processing Effort. *Ph.D. thesis* . University of Edinburgh.

Mohammad, S., & Hirst, G. (2006). Distributional Measures of Concept-Distance: A Task-oriented Evaluation. *Proceedings of the 2006 Conference on Empirical Methods in Nataural Language Processing.* Sydney, Australia.

Narayanan, S., & Harabagiu, S. (2004). Question Answering Based on Semantic Structures. *Proceedings of the 20th international conference on Computational Linguistics.* Geneva, Switzerland.

Narayanan, S., & Mohit, B. (2003). Semantic Extraction with Wide-Coverage Lexical Resources. *Companion Volume of the Proceedings of HLT-NAACL 2003.* Alberta, Canada.

Padó, S. (2007). Cross-Lingual Annotation Projection Models for Role-Semantic Information. *Ph.D. thesis* . Saarland University.

Padó, S., & Lapata, M. (2005). Cross-lingual bootstrapping for Semantic Lexicons: The case of FrameNet. *Proceedings of AAAI-05.* Pittsburgh, Pennsylvania.

Padó, S., & Lapata, M. (2007). Dependency-based Construction of Semantic Space Models. *Computational Linguistics , 33 (2).*

Petruck, M. R. (1996). Frame Semantics. In J. Verschueren, J.-O. Östman, J. Blommaert, & C. Bulcaen (Eds.), *Handbook of Pragmatics 1996.* John Benjamins Publishing Company.

Pitel, G. (2008). Cross-lingual labeling of semantic predicates and roles: A low-resource method based on bilingual L(atent) S(emantic) A(nalysis). In H. Boas (Ed.), *Multilingual FrameNets in Computational Lexicography: Methods and Applications.* Mouton de Gruyter.

Poesio, M., Ishikawa, T., Schulte im Walde, S., & Viera, R. (2002). Acquiring Lexical Knowledge for Anaphora Resolution. *Proceedings of the 3rd Conference on Language Resources and Evaluation, IV.* Las Palmas de Gran Canaria, Spain.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice.*

Sahlgren, M. (2006). The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces. *Ph.D. thesis* . Stockholm University.

Sahlgren, M., & Cöster, R. (2004). Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. *Proceedings of the 20th International Conference on Computational Linguistics.* Geneva, Switzerland.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval.* New-York: McGraw-Hill.

Salton, G., Wong, A., & Yang, C. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM , 18* (11), 613-620.

Schulte im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics , 32 (2)*, 154-194.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics , Special Issue on Word Sense Disambiguation*, 97-123.

Schütze, H., & Pedersen, J. O. (1997). A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management , 33* (3), 307-318.

Ye, P., & Baldwin, T. (2006). Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler. *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006).* Sydney, Australia.

Yoshida, S., Yukawa, T., & Kuwabara, K. (2003). Constructing and Examining Personalized Cooccurrence-based Thesauri on Web Pages. *Proceedings of the 12th International World Wide Web Conference.* Budapest, Hungary.

Appendix A: Stop Word Lists

This section lists the stop words and part-of-speech tags as used in our models.

Excluded words (in English):

```
should      may         such        very        just        also
other       now         then        than        could       who
more        all         it          as          you         not
that        for         to
```

Excluded part-of-speech tags (in English):

```
.           ,           (           )           $           :
?           !           -           ;           DD          DD1
DD2         DDQ         DDQ$        DB          DB2         ART
ZZ1         NUM         PPIO1       PPIO2       PPHO1       PPHO2
PPIS1       PPIS2       PPHS1       PPHS2       APP$        CS
CSW         CSA         CST         PREP        IW          C
```

Excluded words (in German):

```
<unknown>
```

Excluded part-of-speech tags (in German):

```
$.          $,          $(          $)          $$          $:
$?          $!          $-          $;          $/          NUM
ART         PPER        PREP        PRF         PRELS       PTKZU
PWS         KOUI        KOUS        PDS
```

Appendix B: Most Similar Frames

This section lists 100 randomly chosen lexical units, the frames they evoke and the most similar frame according to our model.

```
pathetic A (Subject_stimulus, Desirability) => Mental_property
conceal V (Hiding_objects, Eclipse) => Hiding_objects
delta N (Relational_natural_features) => Natural_features
motorway N (Roadways) => Roadways
obstruct V (Hindering, Eclipse) => Arrest
sick A (Medical_conditions, Biological_urge) => Biological_urge
bearing N (Conduct) => Change_tool
insipid A (Chemical-sense_description) => Exertive_force
ankle N (Observable_bodyparts) => Shaped_part
utilise V (Using) => Using
youth N (People_by_age) => Aggregate
pro N (Expertise) => Competition
slosh V (Self_motion) => Mass_motion
effect N (Objective_influence, Being_in_effect, Subjective_influence)
                                          => Subjective_influence
liquidate V (Killing) => Predicting
boo V (Judgment) => Sounds
sombrero N (Accoutrements) => Accoutrements
botch V (Bungling) => Hit_target
innocent A (Guilt_or_innocence) => Guilt_or_innocence
room V (Residence) => Becoming_aware
spear N (Weapon) => Bearing_arms
electrical A (power N) => Electricity
seamount N (Natural_features) => Prison
sarcophagus N (Containers) => Buildings
gat N (Weapon) => Part_edge
wellington N (Clothing) => Commerce_sell
toxin N (Toxic_substance) => Response
exasperating A (Subject_stimulus) => Typicality
trammel V (Hindering) => Change_tool
surprising A (Subject_stimulus) => Disembarking
summons V (Arrest) => Dimension
dread N (Experiencer_subj) => Experiencer_subj
undergarment N (Clothing) => Clothing
pray V (Rite) => Remembering_information
make N (Type) => Building
forecast V (Predicting) => Change_position_on_a_scale
embellish V (Filling) => Accoutrements
disagreement N (Quarreling) => Quarreling
irritating A (Subject_stimulus) => Experiencer_obj
dampen V (Cause_to_be_wet) => Institutionalization
snarl N (Facial_expression, Sounds) => Perception_active
```

```
sadden V (Experiencer_obj) => Measure_linear_extent
poach V (Apply_heat) => Fleeing
cluck V (Communication_noise) => Motion_noise
indicative A (Sign) => Type
saunter V (Self_motion) => Traversing
forgetful A (Mental_property) => Motion_directional
ruling N (Verdict, Documents) => Organization
chemical A (weapon N) => Weapon
annoy V (Experiencer_obj) => Experiencer_subj
whiff N (Sensation) => Sensation
wet V (Cause_to_be_wet) => Grooming
refuse V (Agree_or_refuse_to_act) => Agree_or_refuse_to_act
tablespoon N (Measure_volume) => Food
elegy N (Text) => Reading
scrimp V (Frugality) => Part_piece
stenosis N (Medical_conditions) => Death
obese A (Body_description_holistic) => Cause_change_of_phase
invention N (Invention, Invention, Achieving_first)
                                          => Jury_deliberation
intimidate V (Experiencer_obj) => Sociability
third A (Ordinal_numbers) => Surpassing
embarrassing A (Subject_stimulus) => Forging
befog V (Eclipse) => Resolve_problem
act V (Performers_and_roles, Conduct, Intentionally_act) => Conduct
bottom A (Part_orientational) => Hostile_encounter
do V (duty N) => Take_place_of
post V (Sending) => Sending
network N (Network) => Network
smuggling N (Smuggling) => Smuggling
recuperation N (Recovery) => Losing_it
system N (Set_of_interrelated_entities, System, Gizmo) => Gizmo
smear V (Placing, Filling) => Apply_heat
terrorize V (Cause_to_experience) => Setting_out
bullet N (Ammunition) => Cause_harm
sicken V (Experiencer_obj) => Sensation
family N (name N) => Being_named
pasta N (Food) => Food
mass N (Quantity, Rite) => Quantity
slip V (Undressing) => Manipulation
distance N (Range) => Shapes
rival V (Evaluative_comparison) => Cause_to_amalgamate
circle N (Shapes, Aggregate) => Natural_features
remedy N (Cure) => Cure
conspiracy N (Offenses, Collaboration) => Killing
cackle V (Communication_noise, Make_noise) => Communication_noise
rid V (Emptying) => Evoking
focus V (Place_weight_on) => Redirecting
frighten V (Experiencer_obj) => Experiencer_subj
hardback N (Text) => Text
convey V (Bringing, Successfully_communicate_message) => Giving
shush V (Become_silent, Silencing) => Misdeed
forge V (Forging) => Forging
```

```
reward N (Rewards_and_punishments) => Judgment
carve V (Create_representation) => Sharpness
crisscross V (Traversing, Path_shape) => Aesthetics
screenplay N (Text) => Performers
big-boned A (Body_description_holistic) => Facial_expression
prejudge V (Partiality) => Feigning
jewelry N (Accoutrements) => Medical_conditions
wrangling N (Quarreling, Hostile_encounter) => Assistance
```