

Towards a Vector Space Model for FrameNet-like Resources

Marco Pennacchiotti*, Diego De Cao[†], Paolo Marocco[†], Roberto Basili[†]

*Computational Linguistics
Saarland University
Saarbrücken, Germany
pennacchiotti@coli-uni.sb.de

[†]DISP
University of Rome Tor Vergata
Rome, Italy
{decao,basili,marocco}@info.uniroma2.it

Abstract

In this paper, we present an original framework to model frame semantic resources (namely, FrameNet) using minimal supervision. This framework can be leveraged both to expand an existing FrameNet with new knowledge, and to induce a FrameNet in a new language. Our hypothesis is that a frame semantic resource can be modeled and represented by a suitable semantic space model. The intuition is that semantic spaces are an effective model of the notion of “being characteristic of a frame” for both lexical elements and full sentences. The paper gives two main contributions. First, it shows that our hypothesis is valid and can be successfully implemented. Second, it explores different types of semantic VSMs, outlining which one is more suitable for representing a frame semantic resource. In the paper, VSMs are used for modeling the linguistic core of a frame, the *lexical units*. Indeed, if the hypothesis is verified for these units, the proposed framework has a much wider application.

1. Introduction

In recent years, NLP research has been focusing on complex tasks, such as Recognizing Textual Entailment (RTE) (Dagan et al., 2006; Giampiccolo et al., 2007), requiring a large amount of semantic knowledge. Part of this knowledge lies at the predicate-argument structure level, in between the syntactic and the deep semantic level. Predicate-argument resources (e.g. FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005)) allow to identify meaning-preserving transformations, such as active/passive, verb alternations and nominalizations, which are crucial in inference based tasks, such as RTE. In the contexts of RTE, consider for example the following pair of text (T) - hypothesis (H):

T: The assassination of JFK in 1963 is an unsolved mystery.

H: JFK was killed in 1963.

An RTE system could infer that the text implies the hypothesis ($T \rightarrow H$) by knowing that in FrameNet the predicates *assassination* and *kill* are in the same frame KILLING, and that for this frame the patterns “VICTIM kill in TIME” and “assassination of VICTIM in TIME” represent a meaning preserving transformation.

Several studies (e.g. (Bar-Haim et al., 2005; Clark et al., 2007; Garoufi, 2007)) have shown that large parts of the RTE-challenges corpora can be solved only relying on repositories of semantic knowledge at the predicate-argument structure level, such as FrameNet and PropBank. For example, (Bar-Haim et al., 2005) report that 31% of the RTE-2 positive test set involves paraphrase at the predicate level. Similarly, (Garoufi, 2007) shows that at least 20% of

the positive examples in the RTE-2 test set can be treated by inferences at the frame level (such as nominalizations and argument variations).

Yet, so far systems based on these resources did not achieve significantly better performance than pure syntactic approaches (e.g.(Burchardt and Frank, 2006)). One of the main reasons is that these resources are manually built: then they are highly accurate but often they have a poor coverage on the test collections. Despite this is a critical issue for their exploitation, not many efforts have been paid so far to automatically or semi-automatically expand these resources.

Another limitation of existing predicate-level resources is the limited support to multilinguality. Multilingual knowledge is becoming more and more important in real Natural Language Processing (NLP) applications. Recently, many researches have focused on using English FrameNet to manually adapt the resource to specific languages – e.g. German (Burchardt et al., 2006), Spanish (Subirats and Petruck, 2003). Unlike PropBank, FrameNet is in fact suitable to cross-lingual induction (Boas, 2005), as frames are mostly defined at the conceptual level, thus allowing cross-lingual interpretation. Yet, all these projects consist in manually defining frame linguistic knowledge (e.g. lexical units) in the specific language, and in manually annotating a large corpus, thus requiring a large human effort. While attempts to automate the annotation process are quite promising, the issue of inducing the resource as a whole in a new language is still an open problem.

In this work, we introduce a new framework to model frame-based resources (namely, FrameNet) using minimal supervision. Our general framework can be leveraged to

expand an existing FrameNet with new knowledge as well as to induce a FrameNet in a specific new language, thus offering a contribution to solve the two aforementioned limitations (lack of coverage and multilinguality).

Our **basic hypothesis** is that a frame semantic resource can be (at least partially) modeled and represented by a suitable semantic vector space model (VSM). The intuition is that a semantic space is a model able to capture the notion of *frame* (i.e. the property of “*being characteristic of a frame*”) for both lexical elements (lexical units, frame elements, etc.) and full sentences. This can be achieved by representing these elements as distributional vectors in the space – i.e. by using co-occurrence vectors.

In this paper we take the first step towards our long term goal of inducing the FrameNet resource. We here aim to give two main fundamental contributions. First, we want to prove that our basic hypothesis is valid and can be successfully implemented in existing VSM frameworks. Second, we explore and experiment different types of semantic VSMs, in order to understand which one is more suitable for representing a frame semantic resource. To achieve this, we present and discuss different VSMs for modeling the linguistic *core* of a frame, the *lexical units* (LU). Indeed, if the hypothesis is verified for these units, we can plan to apply our framework more widely.

The paper is organized as follows. In Section 2. we shortly review the notion of vector space models and report major related researches in that area. In Section 3. we present our model. In Section 4. we describe our experiments and discuss the results. Finally, in Section 5. we draw final conclusions and outline possible future works.

2. Vector space models in NLP

In this section we introduce the notion of vector space model, and outline relevant related work on the subject and on FrameNet expansion techniques.

2.1. Types of Vector Space Models

Vector space models (VSM) are widely used in NLP for representing the meaning of words or other lexical entities. The basic intuition is that the meaning of a target word is somehow defined by the context in which it appears (*Distributional Hypothesis* (Harris, 1964)). The *context* can be defined in different ways: as the set of words surrounding the target word, as the paragraph in which it appears, the document, and so on.

Vector spaces are used to model this intuition, by collecting statistics about the contexts of a target word within a large corpus. In this way, a target word tw is represented by a vector, whose dimensions are the contexts in which it appears. When contexts are words w in an n -window of the target – i.e. tw co-occurs with the context word if w is within n -tokens on the left or on the right – we talk about *word-based VSM*. When contexts are documents or sentences in which the target appears, we talk about *document-based VSM*. The value of each dimension is given by the co-occurrence value of the target word with the given context. In the simplest case, these values are *counts*, i.e. the number of times that the target and the context co-occur in the corpus. More sophisticated association mea-

asures can be used to compute co-occurrence values. The most widely used are conditional probability ($p(w|tw)$) of the word given the target or pointwise mutual information ($pmi(tw, w)$) between the target and the word.

Computationally, a VSM is represented by a matrix, whose each row describes a target word with columns describing contexts. This matrix is used to calculate the distributional similarity between two targets, by computing the distance of their vectors. Different distances are here used expressing similarity measures, such as the cosine between vectors or their Euclidean Distance.

The *original matrix* representing the space can be reduced in dimensionality by applying Singular Value Decomposition (SVD) (Landauer and Dumais, 1997), a matrix decomposition process that creates an approximation of the original matrix, aiming to capture semantic dependencies between contexts. The original space is replaced by a lower dimensional space M_k , called k -space in which each dimension is a derived concept. The original matrix M is transformed in the product of three new matrices: U , S , and V such that $M = USV^T$. Truncating M to its first k dimensions means neglecting the least meaningful dimensions according to the original distribution. M_k captures the same statistical information in a new k -dimensional space, where each dimension is a linear combination of some original features. These newly derived features may be thought of as artificial concepts, each one representing an emerging meaning component as a linear combination of many different words (or contexts).

SVD is widely applied in Latent Semantic Analysis (LSA) for Information Retrieval (Landauer and Dumais, 1997) and usually improves similarity computation. This can be explained by three different reasons. First, SVD tends to remove the random noise, that is implicitly spread in the original matrix and biases the similarity computations. Second, SVD allows to discover the latent meaning of targets, and to compute second-order relations among targets, thus producing more expressive similarity outcomes. Third, similarities are computed within a lower-dimensional space, resulting in a computational speed-up.

2.2. Related Work

Semantic space and the distributional hypothesis have been widely and successfully applied to different language related tasks, such as information retrieval (Salton et al., 1975), harvesting thesauri (Lin, 1998) and paraphrase repositories (Lin and Pantel, 2001). A rich survey is discussed in (Weeds, 2007).

As regards frame semantic resource expansion, many researches have focused on expanding FrameNet knowledge using semi-supervised method. Most of these deal with semantic role labeling, i.e. annotate raw text with frame knowledge. Along the seminal work of (Gildea and Jurafsky, 2002), many machine learning approaches have been proposed, achieving good performance, in strictly supervised settings. Recently, more weakly supervised methods are being explored, using *automatic data expansion* techniques, i.e. leveraging existing annotations to generate new annotations for similar unseen predicates. (Gordon and Swanson, 2007) show that the approach is applicable for

syntactically similar verbs. However, their method requires at least one annotated instance of each new predicate, which limits its practical usefulness.

Other researches have focused on manually creating frame annotated corpora for languages different from English, these including German (Burchardt et al., 2006), Spanish (Subirats and Petruck, 2003), Japanese (Ohara et al., 2004) and French (Pitel, 2006). Recent works propose to semi-automate this long and costly manual process, by using annotation projection techniques on parallel corpora (Yarowsky et al., 2001; Pado and Lapata, 2007).

Yet, so far, to our knowledge there has been no attempt to induce FrameNet knowledge using VSMs, a part from some partial exploratory studies on using LSA spaces for FrameNet induction reported in (Pitel, 1998).

3. Vector Space Models for LUs FrameNet modelling

Using VSM for modeling a FrameNet-like resource poses a fundamental question. What is the relation between the geometry of a vector space model and the linguistic notion of frame? How is the distance in the space correlated with the notion of similarity between frame predicates?

As defined in (Fillmore, 1985), a frame is a conceptual structure modeling a prototypical situation. A frame is evoked in texts through the occurrence of its lexical units (LU). A *lexical unit* is a predicate (a noun, a verb, an adjective, etc.) that linguistically expresses the situation of the frame. LUs of the same frame share semantic arguments. For example the frame KILLING (Figure 1) has LUs such as: *assassin, assassinate, blood-bath, fatal, massacre, murderer, kill, killer, killing, silence, starve, suicide*. These LUs share semantic arguments such as KILLER, INSTRUMENT, CAUSE, VICTIM. In FrameNet, a frame is enriched by a corpus of sentences referring to the situation, in which LUs and their role fillers are individually annotated.

Our far reaching goal is to answer to the above question, and eventually find a way to fully represent a frame in a semantic space. In this work, we focus on the first step, aiming at finding an effective space model for the LUs that represent a frame. The goal is then to have similar vectors for LUs belonging to the same frame (or to related frames). For example, we would expect the lexical units *killer* and *suicide* to be closer in the space, as they both evoke the KILLING frame, with respect to *killer* and *eat*, as these latter express unrelated frames (KILLING vs. INGESTION).

The problem is then to find the space that better approximates the notion of frame. Without looking at all numerous possibilities, we focus our attention on the most plausible spaces. Three different type of spaces are most promising to model LUs: word-based, document-based and syntax-based. These latter are similar to word-based VSM, differing on the fact that contexts are not co-occurring words but co-occurring syntactic relations (e.g. *X-VSubj-man* where *X* is the target word). While all these spaces express distributional similarity, from a semantic perspective, they tend to model different types of relations. Even though the nature of these relations and how they are captured by different spaces is still a matter of debate (see (Sahlgren, 2006;

Mohammad and Hirst, 2005) for an in-depth discussion), most recent studies seem to indicate the following.

Syntax-based spaces. Syntax-based spaces are good at modeling *semantic similarity*. Two target words close in the space are likely to be close also in a is-a hierarchy (Budanitsky and Hirst, 2006), i.e. they are synonyms, antonyms, hyperonyms, cousins, etc. (e.g. *human/man, dog/animal, good/bad*). This is explained by the fact that contexts are syntactic relations, and then targets with the same Part of Speech are much closer than targets of different types. Experiments in (Mohammad and Hirst, 2005; Pado, 2007) support this claim. In other terms, syntax-based spaces tend to capture paradigmatic relations and to disregard syntagmatic relations. According to Saussure, *paradigmatic relations* relate two words likely to appear in the same context but not at the same time (*in absentia*). Syntagmatic relations stand between two words when they are likely to be combined together in the same texts (*in presentia*).

Word-based spaces. These spaces model a more generic notion of *semantic relatedness*. Two targets close in the space are likely to be related by some type of generic semantic relation. The nature of relations captured by these spaces is matter of debate. In practice, word-based spaces have shown effective both to capture syntagmatic and paradigmatic relations. This is supported by the fact that words with similar co-occurrences can be both words occurring together in a text (e.g. *doctor* and *patient* in “*the doctor operated the patient in hospital*”, sharing the same contexts “*operated*”) and substitutional words (e.g. *doctor* and *surgeon* in “*the (doctor—surgeon) operated the patient in hospital*”). Experiments in (Pado, 2007) support this idea, showing that word-based spaces capture syntagmatic relations such as meronymy (*door/house*), conceptual association (*doctor/hospital*) and phrasal association (*private/property*), better than syntax-based spaces, while still capturing paradigmatic relations. The same conclusion is drawn in (Sahlgren, 2006), where word-based space are demonstrated to be highly correlated with a thesaurus containing both syntagmatic and paradigmatic relations.

Document-based spaces. Document spaces are historically used to model *topic similarity*, in the sense that words similar in the space tend to refer to the same topics. This definition has a strong Information Retrieval flavor: words close in the space are those occurring in the same documents, i.e. those focusing on a specific topic. This clearly depends on the nature of the corpus at hand, and on the choice of the context (document, paragraph or sentence). Topic similarity mainly involves co-occurring words (e.g. *doctor/hospital* for the medical topic). Document-based spaces should then better capture syntagmatic relations. Notwithstanding, experimental evidence in (Sahlgren, 2006) have shown that these spaces capture both paradigmatic and syntagmatic relations, as it happens in word-based spaces.

From our perspective, the notion of frame has both a syntagmatic and paradigmatic flavor. Indeed, prototypical situations (i.e. frames) involve different types of participants and facts that in the real world can either stage together

Frame: KILLING		
A KILLER or CAUSE causes the death of the VICTIM.		
Frame Elements	CAUSE	The <u>rockslide</u> <u>killed</u> nearly half of the climbers.
	INSTRUMENT	It's difficult to <u>suicide</u> <u>with only a pocketknife</u> .
	KILLER	<u>John</u> <u>drowned</u> Martha.
	MEANS	The flood <u>exterminated</u> the rats <u>by cutting off access to food</u> .
MEDIUM	John <u>drowned</u> <u>Martha</u> .	
Predicates	annihilate.v, annihilation.n, asphyxiate.v, assassin.n, assassinate.v, assassination.n, behead.v, beheading.n, blood-bath.n, butcher.v, butchery.n, carnage.n, crucifixion.n, crucify.v, deadly.a, decapitate.v, decapitation.n, destroy.v, dispatch.v, drown.v, eliminate.v, euthanasia.n, euthanize.v, . . .	

Table 1: Frame *Killing* from the FrameNet Database.

or be one the substitute of the another (the victim and the killer in the first case, the suicide and the killer in the second case). In the same way, lexicalizations of participants and facts can occur together in the text describing the situation or substitute one another. Then, in our particular setting, LUs of the same frame can be either in a paradigmatic hierarchical relation (e.g. *killer/murderer, suicide/killing*) or in much looser syntagmatic relation (e.g. *assassinate/killer*). In our study we focus on word-based and document-based spaces, as they seem to better capture this ambivalent notion. In the experimental section, we compare the two types of spaces, in order to understand how they express out target notion of frame for LUs. We use both the original spaces and reduced spaces using SVD, as the reduction step should exploit all the advantages described in Section 2..

4. Experiments

4.1. Experiment setup

The goal of our experiment is both to verify the feasibility of the initial hypothesis and to compare the usefulness of the word-based and document-based semantic spaces (SVD-reduced or not). Given a semantic space, the experimental task consists in measuring the similarity between pairs of lexical units extracted from two balanced sets: a *true set* where lexical units are selected to express the same frame (e.g. *killing - suicide, adult - infant*), and a *control set* of lexical units belonging to different frames (e.g. *killing - eat*). Ideally, we would like the similarity measure to rank higher pairs in the true set, and lower pairs in the control set. This would mean that the representation of the lexical units in the semantic space is actually modeling the notion of *frameness*.

As *true set TS* we use all pairs of lexical units belonging to the same frame in FrameNet 2.0. Given the list of about 10,196 lexical units organized in 795 frames, we obtained a *TS* of 147,070 pairs¹. The *control set CS* is created by randomly selecting an equal number of pairs out from the true set. The random choice of the control set simulates the

situation we would face in real tasks, such as selecting the best frame to assign to a new lexical unit. However, in a real setting the algorithm would have to deal with a much higher ratio of false positives than the 50% random pairs.

We evaluate the different spaces using Receiver Operating Characteristic - ROC analysis (Green and Swets, 1996), mixing Sensitivity and Specificity. Given a threshold t applied to the similarity measure S , Sensitivity $Se(t)$ represents the probability of accepting pairs in the true set at threshold t . Specificity $Sp(t)$ represents the probability of rejecting pairs in the false set at t . The *ROC curve* plots $Se(t)$ and $1 - Sp(t)$ at all levels of t . Better methods will have ROC curves more similar to the step function. As evaluation indicators we use AROC and Best Accuracy. *AROC* is the total area under the ROC curve. Statistically, it represents the probability that a method will rank a pair in the true set higher than a randomly chosen pair in the control set. *Best Accuracy* is the highest accuracy among the accuracies at all threshold levels. Higher levels of AROC and Best Accuracy denote more accurate methods.

Counts for building the semantic spaces are computed over the TREC-2002 Vol.2 corpus, consisting of about 110 million words, in 230,401 documents. As similarity measure we chose *cosine similarity*, which is generally acknowledged as a reliable measure over vector spaces.²

4.2. Compared Spaces

Semantic spaces are compared against two baselines: a *simple chance* baseline (**SB**) ranking randomly pairs in *TS* and *CS*, and a *hard baseline* (**HB**) computing as ranking the pointwise mutual information between two LUs across the corpus documents. We compare the following spaces:

OrgWord : original word-based space, without SVD reduction. In this space, contexts are all words occurring more than 30 times in the corpus, filtered using a list of 564 generic stop words. In all we obtained 49,890 context words for the *OrgWord* space. Vectors represent individual target LUs and co-occurrence values are computed within text windows centered

¹The 10,196 LUs comprehend multiple instantiation of a LU in different frames. This count is then reduced to 9,790 unique LUs. We further reduce our initial set discarding multiword LUs and LUs that do not appear in the corpus. The final set thus consists of 6,407 LUs.

²Some studies, such as (Weeds et al., 2004), argue that other measures can emphasize different characteristics of the space, but this aspect is out of the scope of the present work.

SYSTEM	PARAMETERS	AROC	BEST ACCURACY
<i>OrgWord</i>	measure=cpmi , win size=5	0.56	0.55
	measure=cpmi , win size=10	0.55	0.54
	measure=pmi , win size=5	0.56	0.53
	measure=pmi , win size=10	0.54	0.53
	measure=cond.prob. , win size=5	0.55	0.54
	measure=cond.prob. , win size=10	0.55	0.54
<i>LsaWord</i>	measure=cpmi , win size=5 , SVD dim=50	0.68	0.64
	measure=cpmi , win size=5 , SVD dim=100	0.68	0.63
	measure=cpmi , win size=5 , SVD dim=150	0.68	0.63
	measure=cpmi , win size=10 , SVD dim=50	0.66	0.62
	measure=cpmi , win size=10 , SVD dim=100	0.66	0.62
	measure=cpmi , win size=10 , SVD dim=150	0.66	0.62
	measure=pmi , win size=5 , SVD dim=50	0.67	0.63
	measure=pmi , win size=10 , SVD dim=50	0.66	0.62
	measure=cond.prob. , win size=5 , SVD dim=50	0.65	0.61
	measure=cond.prob. , win size=10 , SVD dim=50	0.62	0.58
<i>LsaDoc</i>	SVD dim=100	0.65	0.61
<i>SB</i>		0.50	0.50
<i>HB</i>		0.56	0.57

Table 2: Performance of best performing spaces. Association measure for *Org* and *LsaWord* is *cpmi*. SVD dimension for *LsaWord* and *LsaDoc* is 100.

around the targets. Three different association measures are used: conditional probability, *pmi*, and corrected *pmi* (*cpmi*) according to the formula in (Pantel and Ravichandran, 2004). Two different window sizes have been applied in the experiments: 5 and 10.

LsaWord : reduced spaces obtained by applying SVD reduction onto the word-based spaces. Different *LsaWord* spaces have been created from the *OrgWord* spaces, by applying SVD with the following dimensionality cut k : 50, 100, 150.

LsaDoc : document-based spaces obtained by applying SVD reduction on a document-based one. Contexts for the document-based spaces are the 230,401 documents on the corpus³. Vector components are computed using the *tf*idf* between target LUs and individual documents. We reduce the original spaces by applying SVD with dimensions k set to 50, 100, 150.

4.3. Experiment results

Table 2 reports the AROC and best accuracy results of different spaces. Results show that all spaces outperform the random baseline. In general, semantic spaces seem to be partially successful in capturing the notion of frame. All LSA spaces significantly outperform also the hard baseline, while the non-reduced ones have performance close to it. The fact that LSA spaces perform much better than the original both in AROC and accuracy, indicates that the SVD reduction is successful in neglecting irrelevant information (noise) in the original matrix, and to discover latent meaningful features.

³We also used as sentences as contexts, but results are lower than in the case of documents.

Regarding the comparison between word-based and document-based spaces, the former outperforms the latter of 3%. This seems to suggest that the notion of frame is slightly more tied to paradigmatic relations that syntagmatic, i.e. to words which are close in a is-a hierarchy. This would intuitively confirm the fact that the LUs of frames are in many cases synonyms, antonyms or in a is-a relation. To clarify on this issue, we investigated the correlation between the ranking produced by *LsaWord* and *LsaDocs*, to understand if the two spaces capture or not different aspects of the frame notion. We then run a Spearman’s correlation test, obtaining an overall correlation $\rho = 0.448$. This indicates that the two spaces are significantly correlated, i.e. they have similar rankings. Moreover, they tend to capture the same semantic relations. A look at the best 20 ranked pairs for the two approaches in Table 3 seems to confirm this observation: both spaces promote pairs which are in some paradigmatic relation (e.g. *sleet/shower* and *niece/stepfather*), while syntagmatic relations are less common. An explanation for this similar behavior is also given by the impact of SVD on the semantic modelling: it seems that the pseudo concepts implicit in the dimensionality reduction of LSA better promote paradigmatic relations. This confirms previous studies on the application of an LSA-based method for word sense disambiguation (Basili et al., 2006) where predominant senses were better captured over latent semantic spaces.

As for the different parameterizations, there are no significant differences. In general, windows size of 5 words seem slightly better than those of size 10. Also, *cpmi* appears to be the best association measure, slightly outperforming *pmi* and largely outperforming conditional probability, in all settings. For LSA spaces, the best dimensionality is 50, while 100 and 150 perform comparably.

It is also interesting to notice that all spaces have a very high

ORGWORD	LSAWORD	LSADOCs
segregate, split	granddaughter, father-in-law	snow, sleet
salty, reek	hypertension, ulcer	snow, snowfall
stinging, enjoyable	hepatitis, hypertension	snowfall, sleet
rotten, super	squabble, wrangling	* plateau, sleet
bachelor, court	niece, stepfather	snow, drizzle
sacristy, study	diabetes, hypertension	kidnap, kidnapper
abaya, two-piece	asthma, hypertension	abduct, abduction
generous, miserly	father-in-law, daughter-in-law	snow, shower
design, invent	growl, shriek	sleet, shower
crevasse, bayou	pistol, revolver	drizzle, sleet
partner, conspiracy	diabetes, ulcer	expel, expulsion
skirmish, fistfight	influenza, tuberculosis	shotgun, revolver
prevent, stop	diarrhea, malaria	convict, acquit
country, province	freeway, roadway	shower, drizzle
fascinating, mind-numbing	snowfall, sleet	drizzle, shower
disconcert, bore	bronchitis, cirrhosis	drizzle, scatter
pacify, sober	diabetes, hepatitis	gunfight, shootout
fizzle, beep	aunt, niece	shower, scatter
abaya, one-piece	granddaughter, daughter-in-law	raid, bombing
AK-47, explosive	noodle, biscuit	bombing, raid

Table 3: Best 20 ranked pairs for the best parameterizations of the different spaces.

precision in the first part of the rank (in most cases over 0.90 in the first 10,000 ranked pairs), while in the middle section positive and negative examples mix. This suggests that the spaces can easily capture some basic aspect of the frame notion (leading to high precision in the upper ranking) but at the same time are not able to distinguish other aspects (leading to overlap in the center of the ranking). However, such a high precision in the upper rank allows to directly leverage our method for inferring new LUs, or LUs in languages other than English.

5. Conclusions and Future Work

In this paper we presented an original framework to model FrameNet using minimal supervision, based on semantic spaces. Results of our experiments prove that distributional semantics is at a large extent valuable to model the notion of frame. Results also show that word- and document-based spaces capture similar properties, i.e. they both tend to promote LU pairs according to some paradigmatic relations. This is somehow the effect of the SVD projection that successfully applies to the original spaces: second order relations among LUs are thus effective as models of framedness. As our initial hypothesis has been demonstrated to be valuable, we aim to further explore and leverage our model. First, the promising experimental results show that there is room for the exploration of more complex spaces, able to capture the notion of frame more exhaustively. In this direction we will explore weak syntax-based contexts (e.g. *word+POS*) and try to integrate different spaces. Secondly, we will apply the model to more sophisticated and application-oriented tasks, such as the induction of new LUs from raw texts, the modeling of other frame properties, and the automatic acquisition of example sentences for frames. It is important to emphasize here that the presented model is largely applicable the acquire frame semantics for

languages other than English, and paves the way to a practical architecture for semi-supervised FrameNets construction.

Acknowledgements

This work has partly been funded by the German Research Foundation DFG (grant PI 154/9-3).

6. References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.
- Roy Bar-Haim, Idan Szpektor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Definition and Analysis of Intermediate Entailment Levels*, Ann Arbor, Michigan.
- Roberto Basili, Marco Cammisa, and Alfio Gliozzo. 2006. Integrating domain and paradigmatic similarity for unsupervised sense tagging. In *17th European Conference on Artificial Intelligence (ECAI06)*, Riva del Garda, Italy.
- H. C. Boas. 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478.
- Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.
- Aljoscha Burchardt and Anette Frank. 2006. Approximating Textual Entailment with LFG and FrameNet Frames. In *Proceedings of PASCAL RTE2 Workshop*.
- Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pad, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, Italy.

- Peter Clark, Phil Harrison, John Thompson, William Murray, Jerry Hobbs, and Christiane Fellbaum. 2007. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 54–59, Prague, June. Association for Computational Linguistics.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Quionero-Candela et al., editor, *LNAI 3944: MLCW 2005*, pages 177–190, Milan, Italy. Springer-Verlag.
- Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 4(2):222–254.
- K. Garoufi. 2007. Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. *M.Sc. thesis*, saarland university.
- Daniilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.
- Andrew Gordon and Reid Swanson. 2007. Generalizing semantic role annotations across syntactically similar verbs. In *Proceedings of ACL*, pages 192–199.
- David M. Green and John A. Swets. 1996. *Signal Detection Theory and Psychophysics*. John Wiley and Sons, New York, USA.
- Zellig Harris. 1964. Distributional structure. In Jerrold J. Katz and Jerry A. Fodor, editors, *The Philosophy of Linguistics*, New York. Oxford University Press.
- Tom Landauer and Sue Dumais. 1997. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.
- Dekang Lin and Patrick Pantel. 2001. DIRT-discovery of inference rules from text. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, San Francisco, CA.
- Dekang Lin. 1998. Automatic retrieval and clustering of similar word. In *Proceedings of the Joint International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, Montreal, Canada.
- Saif Mohammad and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. *Submitted for publication*.
- K. Ohara, S. Fujii, T. Ohori, R. Suzuki, H. Saito, and S. Ishizaki. 2004. The Japanese FrameNet project: An introduction. In *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora at LREC 2004*.
- Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2).
- Sebastian Pado, 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Saarland University.
- M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).
- Patrick Pantel and Depak Ravichandran. 2004. Automatically labeling semantic classes. In *Proceedings of HLT-NAACL-04*, Boston, MA.
- Guillaume Pitel, 1998. *Cross-lingual labeling of semantic predicates and roles: A low-resource method based on bilingual L(atent) S(ematic) A(nalysis)*. Mouton De Gruyter,, Berlin/New York.
- Guillaume Pitel. 2006. Using bilingual lsa for framenet annotation of french text from generic resources. In *Workshop on Multilingual Semantic Annotation: Theory and Applications*, Saarbrücken, Germany.
- Magnus Sahlgren. 2006. *The Word-Space Model*. Department of Linguistics, Stockholm University.
- G. Salton, A. Wong, and C. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620.
- C. Subirats and M. Petrucci. 2003. Surprise! Spanish FrameNet! In *Proceedings of the Workshop on Frame Semantics at the XVII. International Congress of Linguists*, Prague.
- J. Weeds, D. Weir, and D. McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of CoLing 2004*, Geneva, Switzerland.
- Julie Weeds. 2007. Measures and applications of lexical distributional similarity. *PhD thesis*.
- David Yarowsky, Grace Ngai, and Roger Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st Human Language Technology Conference*, San Francisco, CA.