Pattern-based Distinction of Paradigmatic Relations for German Nouns, Verbs, Adjectives

Sabine Schulte im Walde and Maximilian Köper

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany

Abstract. This paper implements a simple vector space model relying on lexico-syntactic patterns to distinguish between the paradigmatic relations *synonymy, antonymy* and *hypernymy*. Our study is performed across word classes, and models the lexical relations between German nouns, verbs and adjectives. Applying *nearest-centroid classification* to the relation vectors, we achieve a precision of 59.80%, which significantly outperforms the majority baseline (χ^2, $p<0.05$). The best results rely on large-scale, noisy patterns, without significant improvements from various pattern generalisations and reliability filters. Analysing the classification shows that (i) antonym/synonym distinction is performed significantly better than synonym/hypernym distinction, and (ii) that paradigmatic relations between verbs are more difficult to predict than paradigmatic relations between nouns or adjectives.

1 Introduction

Paradigmatic relations (such as synonymy, antonymy and hypernymy, cf. [1]), are notoriously difficult to distinguish because the first-order co-occurrence distributions of the related words tend to be very similar across the relations. For example, with regard to the sentence *The boy/girl/person loves/hates the cat*, the nominal co-hyponyms *boy, girl* and their hypernym *person* as well as the verbal antonyms *love* and *hate* occur in identical contexts, respectively. Accordingly, while there is a rich tradition on identifying paradigmatically related word pairs in isolation (cf. [2–4] on synonymy, [5–7] on antonymy and [8–10] on hypernymy, among many others), there is little work that has addressed the distinction between two or more paradigmatic relations (such as [11–13] on distinguishing synonyms from antonyms).

The current study applies a simple vector space model to the distinction of paradigmatic relations in German, across the three word classes of nouns, verbs and adjectives. The vector space model is generated in the tradition of lexico-syntactic patterns: we rely on the linear sequences between two simplex words (representing synonyms, antonyms or hypernyms) as vector features in order to predict the lexical semantic relation between the two words. Our hope is that the vector space models using such patterns will unveil differences between the semantic relation pairs. For example, intuitively 'und' (*and*) should be a 1-word pattern to connect synonyms rather than antonyms, while 'oder' (*or*) should be a 1-word pattern to connect antonyms rather than synonyms. The

pattern-based approach to distinguish lexical semantic relations has first been proposed by [8] to identify *noun hypernyms*; subsequent prominent pattern-based approaches are [14, 15] who identified *noun meronyms*; [16] on *noun causality*; [17] on *verb similarity, strength, antonymy, enablement, happens-before*; [18] on *noun hypernymy, meronymy, succession, reaction, production*; and [19] on *noun relational analogies*. (See Section 2 for more details on related work.) Our main questions with regard to the study can be summarised as follows.

- Can lexico-syntactic patterns distinguish between paradigmatic relations?
- Which relations are more difficult to distinguish than others?
- What are the differences across word classes?

2 Related Work

Although there are not many approaches in Computational Linguistics that explicitely addressed the distinction of paradigmatic semantic relations, there is a rich tradition on either synonyms or antonyms or hypernyms. Prominent work on identifying ***synonyms*** has been provided by Edmonds who employed a co-occurrence network and second-order co-occurrence (e.g., [20–22, 2]), and Curran who explored word-based and syntax-based co-occurrence for thesaurus construction (e.g., [23, 3]). [24] presented two methods (using patterns vs. bilingual dictionaries) to identify synonyms among distributionally similar words; [4] compared a standard distributional approach against cross-lingual alignment; [25] defined a vector space model for word meaning in context, to identify synonyms and the substitutability of verbs. Most computational work addressing ***hypernyms*** was performed for nouns, cf. the lexico-syntactic patterns by [8] and an extension of the patterns by dependency paths [10]. [26, 27] represent systems that identify hypernyms in distributional spaces. Examples of approaches that addressed the automatic construction of a hypernym hierarchy (for nouns) are [28, 9, 29–31]. Hypernymy between verbs has been addressed by [32–34]. Comparably few approaches have worked on the automatic induction of ***antonyms***. A cluster of approaches in the early 90s tested the co-occurrence hypothesis, e.g., [35, 36, 5]. In recent years there have been approaches to antonymy that were driven by text understanding efforts, or being embedded in a larger framework to identify contradiction [37, 6, 7, 38].

Among the few approaches that distinguished *between* paradigmatic semantic relations we only know about systems addressing ***synonyms vs. antonyms***. [24] implemented a similarity measure to retrieve distributionally similar words for constructing a thesaurus. They used a post-processing step to filter out any words that appeared with the patterns 'from X to Y' or 'either X or Y' significantly often, as these patterns usually indicate opposition rather than synonymy. [11] tackled the task within a pattern-based approach (see below). A recent study by [13], whose main focus was on the identification and ranking of opposites, also discussed the task of synonym/antonym distinction as a specific application of their findings.

Regarding pattern-based approaches to identify and distinguish lexical semantic relations in more general terms, [8] was the first to propose lexico-syntactic patterns as empirical pointers towards relation instances. Her goal was to identify pairs of nouns where one of the nouns represented the hypernym of the other. She started out with a handful of manual patterns such as

NP_i {, NP_j}* {,} *and other* NP_k

that were clear indicators of the lexical relationship (in this case with NP_i and NP_j representing hyponyms of NP_k), and used bootstrapping to alternately (i) find salient instances on the basis of the patterns, and (ii) rely on the enlarged set of pair instances to identify more salient patterns that are indicators of the relationship. Hearst demonstrated the success of her approach by comparing the retrieved noun pairs with WordNet lexical semantic relation pairs.

Girju [16] distinguished pairs of nouns that are in a causal relationship from those that are not. Differently to Hearst, she only relied on a single pattern

NP_i *verb* NP_k

that represented a salient indicator of causation between two nouns (with NP_i representing the cause and NP_k the effect) but at the same time was a very ambiguous pattern. Girju used a Decision Tree on 683 noun pairs and predicted the existence of a causal relation with a precision of 73.91% and a recall of 88.69%; in addition, she applied the causation prediction to question answering and reached a significant improvement. In [15], Girju and colleagues extended the lexical relation work to part–whole relations, applying a supervised, knowledge-intensive approach, mainly relying on WordNet and semantically annotated corpora. As in the earlier work, the task was to distinguish positive and negative relation instances. While they reached an f-score of 82.05%, they noted that many of the lexico-syntactic patterns were highly ambiguous (i.e., depending on the context they indicated different relationships).

[17] were the first to apply pattern-based relation extraction to verbs. For five non-disjoint lexical semantic relations (*similarity, strength, antonymy, enablement, happens-before*) they manually defined patterns and then queried Google to estimate joint pair–pattern frequencies for WordNet pairs as well as verb pairs generated by DIRT [39]. The accuracy for predicting whether a certain pair undergoes a certain semantic relationship varied between 50% and 100%, for relation set sizes of 2–41.

[18] developed *Espresso*, a weakly-supervised system that exploits patterns in large-scale web data. Similarly to [15], they used generic patterns, but relied on a bootstrapping cycle combined with reliability measures, rather than manual knowledge resources. Espresso worked in three phases: pattern induction, pattern selection and instance extraction. Starting with seed instances for the lexical semantic relations, the bootstrapping cycle iteratively induced patterns and new relation instances by web queries. Each induction step was combined with filtering out the least salient patterns/instances by reliability measures. The approach was applied to five noun-noun lexical semantic relations (*hypernymy, meronymy, succession, reaction, production*) and reached accuracy values between 49% and 91%, depending on the data and the relationship.

The work by Turney also includes approaches to extract and distinguish word pairs with regard to their lexical semantic relation. He developed a framework called *Latent Relational Analysis (LRA)* [40, 41, 19] that relied on corpus-based patterns between words in order to model relational similarity, i.e., similarity between word pairs *A:B::C:D* such that *A is related to B as C is related to D*. In his framework, a vector space model was populated with word pairs as the targets and patterns as the pair features. The patterns were derived from web corpora, and the cosine was used to measure the relational similarity between two word pairs. Turney applied a range of modifications to his basic setup, including a step-wise generalisation of the patterns by wild-cards instead of specific word types; extension of target pairs by synonyms to the words within a pair, as determined by Lin's thesaurus [42]; feature reduction by Singular Value Decomposition; etc. LRA has been applied to predict analogies in semantic relation pairs, to classify noun-modifier pairs according to the noun-noun semantic relation; to identify TOEFL synonyms; to answer SAT questions; to distinguish synonyms and antonyms; among others.

3 Paradigmatic Relation Datasets

The dataset of paradigmatic relations used in our research has been collected independently of the specific classification task in this paper. Based on a selection of semantic relation targets across the three word classes nouns, verbs and adjectives, we collected antonyms, synonyms and hypernyms for these targets via crowdsourcing. The following steps describe the creation of the dataset in more detail.

1. **Target source, semantic classes and senses:** We selected GermaNet[1] [43–45] as the source for our semantic relation targets. GermaNet is a lexical-semantic taxonomy for German that defines semantic relations between word senses, in the vein of the English WordNet [46]. Relying on GermaNet version 6.0 and the respective *JAVA API*, we generated lists of all nouns, verbs and adjectives, according to their semantic class (as represented by the file organisation), and also extracted the number of senses for each lexical item.
2. **Target frequencies:** Relying on the German web corpus *sdeWaC* (version 3), we extracted corpus frequencies for all lexical items in the GermaNet files, if available. The *sdeWaC* corpus [47] is a cleaned version of the German web corpus *deWaC* created by the *WaCky* group [48]. It contains approx. 880 million words with lemma and part-of-speech annotations [49] and can be downloaded from http://wacky.sslmit.unibo.it/.
3. **Target selection:** Using a stratified sampling technique, we randomly selected 99 nouns, 99 adjectives and 99 verbs from the GermaNet files. The random selection was balanced for

[1] www.sfs.uni-tuebingen.de/lsd/

(a) the ***size of the semantic classes***,[2] accounting for the 16 semantic adjective classes and the 23 semantic classes for both nouns and verbs;

(b) ***three polysemy classes*** according to the number of GermaNet senses: I) monosemous, II) two senses and III) more than two senses;

(c) ***three frequency classes*** (type frequency in sdeWaC): I) *low* (200–2,999), II) *mid* (3,000–9,999) and III) *high* (≥10,000).

The total number of 99 targets per word class resulted from distinguishing 3 sense classes and 3 frequency classes, $3 \times 3 = 9$ categories, and selecting 11 instances from each category, in proportion to the semantic class sizes.

4. **Semantic relation generation:** An experiment hosted by Amazon Mechanical Turk (AMT)[3] collected synonyms, antonyms and hypernyms for each of our 3×99 targets. For each word class and semantic relation, the targets were distributed randomly over 9 batches including 9 target each. In order to control for spammers, we in addition included two German fake words into each of the batches, in random positions of the batches. If participants did not recognise the fake words, all of their data were rejected. We asked for 10 participants per target and relation, resulting in 3 word classes $\times$ 99 targets $\times$ 3 relations $\times$ 10 participants $= 8,910$ target–response pairs. Table 1 shows some examples of the generated target–response pairs across the word classes and relations. The examples are accompanied by the *strength* of the responses, i.e., the number of participants who provided the response.

Table 1. Examples of target–response pairs across word classes and semantic relations.

	ANT		SYN		HYP	
NOUN	*Bein/Arm* (leg/arm)	10	*Killer/Mörder* (killer)	8	*Ekel/Gefühl* (disgust/feeling)	7
	Zeit/Raum (time/space)	3	*Gerät/Apparat* (device)	3	*Arzt/Beruf* (doctor/profession)	5
VERB	*verbieten/erlauben* (forbid/allow)	10	*üben/trainieren* (practise)	6	*trampeln/gehen* (lumber/walk)	6
	setzen/stehen (sit/stand)	4	*setzen/platzieren* (place)	3	*wehen/bewegen* (wave/move)	3
ADJ	*dunkel/hell* (dark/light)	10	*mild/sanft* (smooth)	9	*grün/farbig* (green/colourful)	5
	heiter/trist (cheerful/sad)	2	*bekannt/vertraut* (familiar)	4	*heiter/hell* (bright/light)	1

We decided in favour of this very specific dataset and against directly using the GermaNet relations, for the following reason. Although GermaNet aims to include examples of all three relation types for each of the three parts-of-speech (nouns, verbs, adjectives), coverage of these can be low in places, as depending on the part-of-speech some semantic relations apply more naturally than others [50]. For example, the predominant semantic relation for nouns is hypernymy, whereas the predominant semantic relation for adjectives is antonymy. As a result, GermaNet does not always provide all three relations with regard to a specific lexical unit.

[2] For example, if an adjective GermaNet class contained a total of 996 word types, and the total number of all adjectives over all semantic classes was 8,582, and with 99 stimuli collected in total, we randomly selected $99 * 996/8,582 = 11$ adjectives from this semantic class.

[3] https://www.mturk.com

4 Experiments

The goal of our experiments was to distinguish between the three paradigmatic relations *antonymy, synonymy, hypernymy*. The following subsections describe the setup of the experiments (Section 4.1) and the results (Section 4.2).

4.1 Setup

Dataset: The experiments rely on a subset of the collected pairs as described in the previous section, containing those target–response pairs that were provided at least twice (to ensure reliability) and without ambiguity[4] between the relations. Table 2 shows the distribution of the target–response pairs across classes and relations. The target–relation pairs were randomly divided into 80% training pairs and 20% test pairs with regard to each class–relation combination.

Table 2. Target–response pairs.

	ANT	SYN	HYP
NOUN	95	90	97
VERB	75	76	74
ADJ	62	62	61

In addition to using this dataset, the overall best experiments were performed on a variant that investigated the influence of polysemy among the targets and responses. We relied on the same dataset but distinguished between monosemous vs. polysemous target–response pairs. I.e., we divided the training pairs and the test pairs into two sets for each class–relation combination, one containing only pairs where both the target and the response were monosemous, and one containing only pairs where either the target or the response was polysemous, according to the definitions in GermaNet. (The third case, that both target and response are polysemous, did not show up in our dataset.)

Patterns: For all our target–response pairs, we extracted the lexico-syntactic patterns between the targets and the responses. The basic patterns (to be refined; see below) relied on raw frequencies of lemmatised patterns. Since hypernymy requires the definition of pattern direction, all our patterns were marked by their directionality. As corpus resource, we relied on WebKo, a predecessor version of the sdeWaC (cf. Section 3), which comprises more data (approx. 1.5 billion words in comparison to 880 million words) but is less clean. We found a total of 95,615/54,911/21,350 pattern types for the nouns/verbs/adjectives, when neither the length of the patterns was restricted or any kind of generalisation applied. The basic patterns were varied as follows.

[4] Ambiguity between the relations arose when the same response was provided for a target with regard to two semantic relations. For example, *Maschine* 'machine' was provided both as a synonym and a hypernym of the noun *Gerät* 'device, machine'. We disregarded such ambiguous cases in this paper.

1. *Morpho-syntactic generalisation:* The patterns were generalised by (i) substituting each common noun, proper name, adjective and determiner by its part-of-speech; (ii) deleting all non-alphabetic characters from the patterns.
2. *Mutual information variants:* We used point-wise mutual information values (pmi) [51, 18] instead of raw pattern frequencies, and implemented two variants: (i) *pmi(relation,pattern)* and (ii) *pmi(pair,pattern)*, thus enforcing the strengths of patterns that were (i) strong indicators of a specific relation or (ii) strong indicators for specific pairs.
3. *Length restriction:* The lengths of the patterns were restricted to maximally 1, 2, ... 100 words between the targets and the responses.
4. *Frequency restriction:* Only patterns with a frequency of at least 1, 2, ... 10, 20, 50, 100 were taken into account, ignoring low-frequent patterns.
5. *Reliability:* The least reliable patterns were deleted from the vector space dimensions. Reliability was determined as in [18]:

$$reliability(pattern) = \frac{\sum_{i \in I}(\frac{pmi(i,pattern)}{max_{pmi}}) \times reliability_i(i)}{|I|} \quad (1)$$

with i representing a pair instance and I the set of all pairs. The value of $reliability_i$ was instantiated by the strength of the pair in our dataset.

Classification and Evaluation: We implemented a simple[5] *nearest-centroid classification* (also known as *Rocchio Classifier* [52]) to distinguish between the paradigmatic relation pairs. For each word class, we calculated three mean vectors, one for each lexical semantic relation (antonymy, synonymy, hypernymy), as based on the training pairs. We then predicted the semantic relation for the test pairs in each word class, by choosing for each test pair the most similar mean vector, as determined by *cosine*.

This 3-way classification to distinguish between the three paradigmatic relations was performed across the various conditions described above, to identify the types and variations of patterns that were most useful. In a follow-up step we applied the most successful condition to 2-way classifications that aimed to distinguish between two paradigmatic relations (antonyms vs. synonyms, antonyms vs. hypernyms, synonyms vs. hypernyms). The 2-way classifications were to provide insight into more or less difficult relation pairings.

All predictions were evaluated by *precision*, the proportion of predictions we made that were correct. Since many variations of the pattern features effected the number of patterns, we also calculated *recall*, the proportion of test pairs for which we could make a prediction based on the vector dimensions. Harmonic *f-score* then helped us to decide about the overall quality of the conditions in relation to each other.

[5] We also applied standard approaches that were relevant to the task, such as Decision Trees and k-Nearest-Neighbour, but our simple approach outperformed them.

4.2 Results

Table 3 shows the results of the pattern-based distinctions in the 3-way relation classification experiments. In the first column the result relies on the basic setup, i.e., using all unaltered patterns as vector features. This *basic* result outperforms the majority baseline (44%) significantly[6] ($p < 0.05$), and is at the same time (a) significantly better than relying on the part-of-speech generalisation ($p < 0.1$), (b) not significantly better than relying on the alphanumeric generalisation and (c) significantly better than the pmi versions of the patterns. Interestingly, optimising the patterns by disregarding very long patterns or disregarding patterns with low frequencies does not improve the basic setup: the best results of these optimisations (see columns *length* and *freq* in Table 3) are exactly the same.

Applying the basic setup to monosemous (*mono*) vs. polysemous (*poly*) relation pairs demonstrates that (a) the lexical semantic relations for monosemous word pairs are easier to predict than for pairs involving polysemy (precision: 64.71 vs. 53.01); (b) the polysemous word pairs activate more pattern types (recall: 45.83 vs. 36.67). Both *mono* and *poly* are significantly better than the baseline ($p < 0.1$).

Table 3. 3-way classification results across conditions.

	Pattern variations							Polysemy	
	basic	generalisations		length	freq	pmi		mono	poly
		pos	alpha			rel,pat	pair,pat		
precision	****59.80**	46.85	52.94	****59.80**	****59.80**	48.04	35.29	*64.71	*53.01
recall	**48.41**	41.27	42.86	**48.41**	**48.41**	38.89	28.57	36.67	45.83
f-score	**53.51**	43.88	47.37	**53.51**	**53.51**	42.98	31.58	46.81	49.16

Figures 1 to 3 show the impact of reducing the number and types of patterns with regard to length, frequency and reliability. Figure 1 demonstrates that reducing the vector space to short patterns of maximally 1, 2, ..., 10 words (i.e., deleting very long and specific patterns that appeared between targets and responses) does almost have no impact on the prediction results. In fact, all patterns seem to provide salient information for the classification, as precision, recall and f-score all monotonically increase when including more and longer patterns. The difference between the best result (including all patterns) and the worst result (including only patterns of length 1) is however not significant. Figure 2 demonstrates that deleting infrequent patterns from the vector space (i.e., deleting patterns with a frequency of less than 2, 3, ..., 100) does also have no strong impact on the prediction results. Even low-frequent patterns seem to provide salient information for the classification, as precision, recall and f-score all monotonically decrease when including only more frequent patterns. Again, the difference between the best result (including all patterns) and the worst result (including only high-frequent patterns) is not significant. Figure 3 shows the effect of deleting x% of the most unreliable patterns, with x = 0.01%, 0.02%,

[6] All significance tests have been performed with χ^2. Significance levels are marked at the precision values with $^*p \leq 0.1$, $^{**}p \leq 0.05$ and $^{***}p \leq 0.01$, if applicable.

Fig. 1. Deleting long patterns.

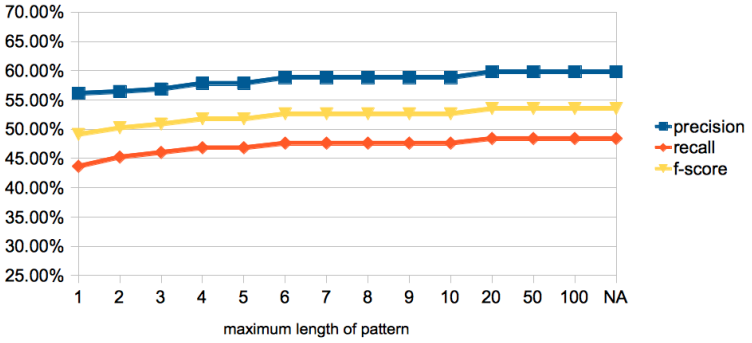


Fig. 2. Deleting infrequent patterns.

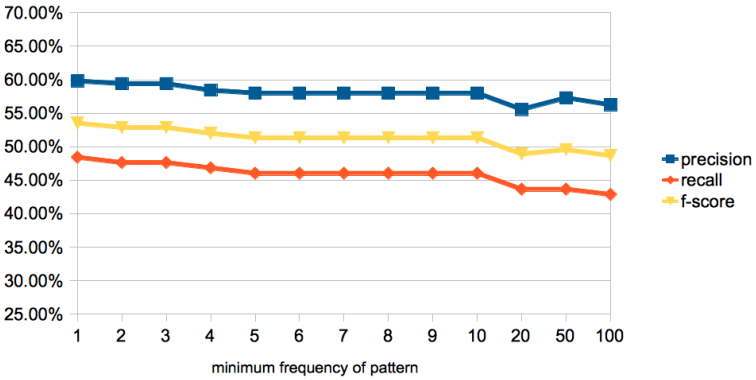
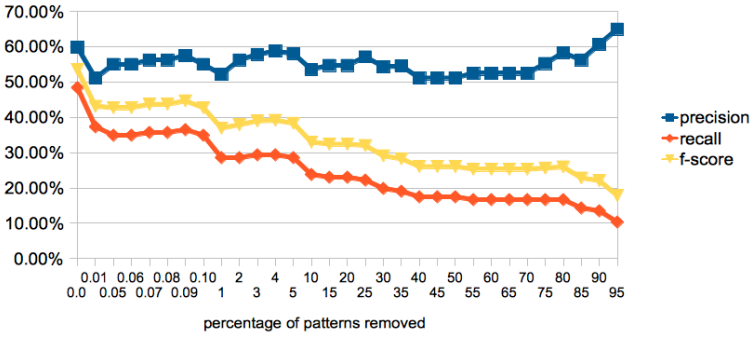


Fig. 3. Deleting unreliable patterns.



..., 1%, 2%, ..., 10%, 15%, 20%, ..., 95%. The plot demonstrates that deleting unreliable patterns does have an impact on the quality of the prediction. Most notably, precision drops severely from 59.80% to 51.09% when deleting the most unreliable patterns, and goes up to 65% when only using the 5-10% most reliable patterns. Recall and f-score monotonically decrease when deleting patterns. Even unreliable patterns seem to provide salient information for the classification. At the same time, we achieved our best precision with 5-10% of the patterns only.

Table 4 shows the results of the pattern-based distinctions in the 2-way relation classification experiments. As the baselines are different,[7] we list them in the table. The table demonstrates that the pair-wise distinction between the relation pairs works differently well for the three type. The antonym/synonym distinction performed best, the synonym/hypernym distinction performed worst. While both the antonym/synonym and the antonym/hypernym distinction are significantly better than the baseline, the synonym/hypernym distinction is not.

Table 4. 2-way classification results.

	ANT/SYN	ANT/HYP	SYN/HYP
baseline	55.00	50.00	55.00
precision	***78.79	**68.06	63.64
recall	64.20	55.68	50.60
f-score	70.75	61.25	56.38

Table 5 shows the confusion matrix for the 2-way relation distinctions, along with the respective precision scores. The *all* column corresponds to the results in Table 4, the other columns distribute these counts over the word classes. Across the three word classes (*all*), the distinction between antonyms and synonyms is significantly better ($p < 0.1$) than the distinction between synonyms and hypernyms. The other differences (ANT/SYN vs. ANT/HYP; ANT/HYP vs. SYN/HYP) are not significant. So the most difficult relation distinction to predict is synonyms vs. hypernyms.

The confusion matrix demonstrates where the incorrect predictions mainly come from: in the antonym/hypernym distinction, half of the hypernyms were predicted as antonyms; in the synonym/hypernym distinction, even more than half of the hypernyms were predicted as synonyms. While there are also other incorrect predictions, these two cases are striking.

Looking at the results with regard to the three word classes, the predictions of verb relations were in all 2-way distinctions worse than those for nouns and adjectives. The differences for verbs vs. nouns on predicting the synonym/hypernym distinction is significant ($p < 0.05$), the other differences are not significant. The noun and adjective relation prediction is similarly good, without remarkable differences, even though one might have expected that the predictions of the 'core' relations (synonymy and hypernymy for nouns; synonymy and antonymy for adjectives) should be better with regard to the respective word class.

[7] Since there are different amounts of antonym/synonym, antonym/hypernym and synonym/hypernym pairs in the final dataset, the majority baseline varies.

Table 5. Confusion matrix (2-way relation distinction).

	NOUN			VERB			ADJ			*all*		
	ANT	SYN	prec	ANT	SYN	prec	ANT	SYN	prec	ANT	SYN	prec
ANT	**15**	2	77.42	**7**	3	70.59	**8**	1	88.89	**30**	6	78.79
SYN	5	**9**		2	**5**		1	**8**		8	**22**	
	ANT	HYP	prec	ANT	HYP	prec	ANT	HYP	prec	ANT	HYP	prec
ANT	**15**	2	74.19	**8**	2	54.55	**8**	1	73.68	**31**	5	68.06
HYP	6	**8**		8	**4**		4	**6**		18	**18**	
	SYN	HYP	prec	SYN	HYP	prec	SYN	HYP	prec	SYN	HYP	prec
SYN	**13**	1	75.00	**5**	2	42.11	**8**	1	68.42	**26**	4	63.64
HYP	6	**8**		9	**3**		5	**5**		20	**16**	

5 Discussion

The results in the previous section demonstrated that a pattern-based vector space model is able to distinguish between paradigmatic relations: The precision of our basic pattern set in the 3-way relation classification (59.80%) significantly outperformed the majority baseline, $p<0.05$. In the 2-way relation classification, the same patterns achieved precision values of 78.79% for antonym/synonym distinction (significant, $p<0.01$), 68.06% for antonym/hypernym distinction (significant, $p<0.05$), and 63.64% for synonym/hypernym distinction (not significant).

None of the variations to the patterns we performed resulted in significant improvements of the basic setup. Even more, generalisations of the patterns by (i) replacing words with their parts-of-speech or by (ii) deleting all non-alphabetic characters made the results worse, in case (i) even significantly ($p < 0.1$). Similarly, the precision results decreased (in some cases even significantly) when we applied mathematical variations and filters to the patterns, by (i) replacing the pattern frequencies by point-wise mutual information scores as well as when (ii) incorporating a filter for unreliable patterns as adopted from [18].

On the one hand, it is not surprising that generalisations of patterns are not successful because it is very difficult to identify –within a large-scale vector space– those aspects of patterns that contribute to subtle distinctions between relation pairs, and those that will not. For example, if we generalise over specific words by their parts-of-speech this might be helpful in some cases (e.g., we find 'und zwei' *(and two)* as well as 'und sieben' *(and seven)*, where we could generalise over the cardinal number) but contra-productive in others (e.g., we find 'Haar und' *(hair and)* as strong indicator for adjective antonyms and 'Land und' *(country and)* as strong indicator for adjective hypernyms, where generalising over nouns would delete the relation-specific distinction). Similarly, generalising over punctuation might be helpful in some cases (e.g., we find 'und d Arme immer' (lemmatised version of 'und die Armen immer') *and the poor always* as well as ', d Arme immer' *, the poor always* as strong indicators for adjective antonyms) but contra-productive in others (e.g., '/' is a strong indicator for adjective antonymy, while '(' is a strong indicator for adjective hypernymy, and ',' is a strong indicator across all adjectival relations).

On the other hand, we would have expected pmi variants to have a positive effect on the prediction strength of the patterns because they should be able to strengthen the contributions of more salient and weaken the contributions of less salient patterns or pairs. Of course, it is possible that our experimental setup does not sufficiently enforce strong features to outplay weak features. In previous work, many of the existing approaches [8, 16, 15, 18] worked within a bootstrapping cycle, i.e., (1) starting with a small set of clearly distinguishing patterns for a small set of prototypical relation instances, (2) increasing the set of relation pairs on the basis of these patterns and large-scale corpus data, (3) using the new pairs to identify new patterns, (4) filtering the patterns for reliability, etc. It was beyond the scope of this study but might be interesting to implement a variant of our setup that incorporates a bootstrapping cycle. However, we would like to emphasise that we doubt that bootstrapping improves our results because our experiments clearly demonstrated that the salient information in the patterns lies within infrequent as well as frequent patterns, and within short as well as long patterns, and within less reliable as well as strongly reliable patterns. This is in accordance with [19] who demonstrated that large-scale and potentially noisy patterns outperform feature vectors with carefully chosen patterns.

It is difficult to numerically compare our results with related work on pattern-based relations because (i) many previous approaches have tried to *identify* semantic relations pairs, rather than distinguish them [8, 16, 17, 15, 18], and (ii) most of the approaches focused on one semantic relation at the same time [8, 16–18]. Concerning (i), our approach is different in that we *distinguish* between relation pairs; we could however also apply our classification to *identify* additional relation pairs, assuming that we first extract a set of candidate pairs. Concerning (ii), our approach is different in that we focus on 2 or 3 semantic relations at the same time, and in addition the distributional differences between paradigmatic relations are subtle (cf. Section 1). With regard to both (i) and (ii), Turney's work is most similar to ours. [11] achieved a precision of 75% on a set of 136 synonym/antonym questions, with a majority class baseline of 65.4%, in comparison to our synonym/antonym distinction achieving 78.79% with a majority baseline of 55%.

6 Conclusion

This paper presented a vector space model relying on lexico-syntactic patterns to distinguish between the paradigmatic relations *synonymy, antonymy* and *hypernymy*. Our best results achieved a precision score of 59.80%, which significantly outperformed the majority baseline. Interestingly, our original noisy patterns performed better than any kind of standard generalisation or reliability filter. We also showed that (i) antonym/synonym distinction is performed significantly better than synonym/hypernym distinction; (ii) paradigmatic relations between verbs are more difficult to predict than paradigmatic relations between nouns or adjectives; and (iii) paradigmatic relations between monosemous words are easier to predict than those involving a polysemous word.

References

1. Murphy, M.L.: Semantic Relations and the Lexicon. Cambridge University Press (2003)
2. Edmonds, P., Hirst, G.: Near-Synonymy and Lexical Choice. Computational Linguistics **28**(2) (2002) 105–144
3. Curran, J.: From Distributional to Semantic Similarity. PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh (2003)
4. van der Plas, L., Tiedemann, J.: Finding Synonyms using Automatic Word Alignment and Measures of Distributional Similarity. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia (2006) 866–873
5. Fellbaum, C.: Co-Occurrence and Antonymy. Lexicography **8**(4) (1995) 281–303
6. Harabagiu, S.M., Hickl, A., Lacatusu, F.: Negation, Contrast and Contradiction in Text Processing. In: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA (2006) 755–762
7. Mohammad, S., Dorr, B., Hirst, G.: Computing Word-Pair Antonymy. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Waikiki, Hawaii (2008) 982–991
8. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France (1992) 539–545
9. Caraballo, S.A.: Automatic Acquisition of a Hypernym-labeled Noun Hierarchy from Text. PhD thesis, Brown University (2001)
10. Snow, R., Jurafsky, D., Ng, A.Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery. Advances in Neural Information Processing Systems **17** (2004) 1297–1304
11. Turney, P.D.: A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK (2008) 905–912
12. Yih, W.T., Zweig, G., Platt, J.C.: Polarity Inducing Latent Semantic Analysis. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea (2012) 1212–1222
13. Mohammad, S.M., Dorr, B.J., Hirst, G., Turney, P.D.: Computing Lexical Contrast. Computational Linguistics **39**(3) (2013) To appear.
14. Berland, M., Charniak, E.: Finding Parts in Very Large Corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland, MD (1999) 57–64
15. Girju, R., Badulescu, A., Moldovan, D.: Automatic Discovery of Part-Whole Relations. Computational Linguistics **32**(1) (2006) 83–135
16. Girju, R.: Automatic Detection of Causal Relations for Question Answering. In: Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering – Machine Learning and Beyond, Sapporo, Japan (2003) 76–83
17. Chklovski, T., Pantel, P.: VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain (2004) 33–40
18. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: Proceedings of the 21st International

Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia (2006) 113–120
19. Turney, P.D.: Similarity of Semantic Relations. Computational Linguistics **32**(3) (2006) 379–416
20. Edmonds, P.: Choosing the Word most typical in Context using a Lexical Co-occurrence Netword. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain (1997) 507–509
21. Edmonds, P.: Translating Near-Synonyms: Possibilities and Preferences in the Interlingua. In: Proceedings of the AMTA/SIG-IL Second Workshop on Interlinguas, Langhorne, PA (1998) 23–30
22. Edmonds, P.: Semantic Representations of Near-Synonyms for Automatic Lexical Choice. PhD thesis, Department of Computer Science, University of Toronto (1999) Published as technical report CSRI-399.
23. Curran, J.: Ensemble Methods for Automatic Thesaurus Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2002) 222–229
24. Lin, D., Zhao, S., Qin, L., Zhou, M.: Identifying Synonyms among Distributionally Similar Words. In: Proceedings of the International Conferences on Artificial Intelligence, Acapulco, Mexico (2003) 1492–1493
25. Erk, K., Padó, S.: A Structured Vector Space Model for Word Meaning in Context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Waikiki, Hawaii (2008) 897–906
26. Weeds, J., Weir, D., McCarthy, D.: Characterising Measures of Lexical Distributional Similarity. In: Proceedings of the 20th International Conference of Computational Linguistics, Geneva, Switzerland (2004) 1015–1021
27. Lenci, A., Benotto, G.: Identifying Hypernyms in Distributional Semantic Spaces. In: Proceedings of the 1st Joint Conference on Lexical and Computational Semantics, Montréal, Canada (2012) 75–79
28. Caraballo, S.A.: Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland, MD (1999) 120–126
29. Velardi, P., Fabriani, P., Missikoff, M.: Using Text Processing Techniques to Automatically enrich a Domain Ontology. In: Proceedings of the International Conference on Formal Ontology in Information Systems, Ogunquit, ME (2001) 270–284
30. Cimiano, P., Schmidt-Thieme, L., Pivk, A., Staab, S.: Learning Taxonomic Relations from Heterogeneous Evidence. In: Proceedings of the ECAI Workshop on Ontology Learning and Population. (2004)
31. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic Taxonomy Induction from Heterogenous Evidence. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia (2006) 801–808
32. Fellbaum, C.: English Verbs as a Semantic Net. Journal of Lexicography **3**(4) (1990) 278–301
33. Fellbaum, C., Chaffin, R.: Some Principles of the Organization of Verbs in the Mental Lexicon. In: Proceedings of the 12th Annual Conference of the Cognitive Science Society of America. (1990) 420–427
34. Fellbaum, C.: A Semantic Network of English Verbs. [46] 69–104
35. Charles, W., Miller, G.: Contexts of Antonymous Adjectives. Applied Psycholinguistics **10** (1989) 357–375
36. Justeson, J.S., Katz, S.M.: Co-Occurrence of Antonymous Adjectives and their Contexts. Computational Linguistics **17** (1991) 1–19

37. Lucerto, C., Pinto, D., Jiménez-Salazar, H.: An Automatic Method to Identify Antonymy Relations. In: Proceedings of the IBERAMIA Workshop on Lexical Resources and the Web for Word Sense Disambiguation, Puebla, Mexico (2004) 105–111
38. de Marneffe, M.C., Rafferty, A.N., Manning, C.D.: Finding Contradictions in Text. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH (2008) 1039–1047
39. Lin, D., Pantel, P.: DIRT – Discovery of Inference Rules from Text. In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, San Francisco, CA (2001) 323–328
40. Turney, P.D.: Measuring Semantic Similarity by Latent Relational Analysis. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland (2005) 1136–1141
41. Turney, P.D.: Expressing Implicit Semantic Relations without Supervision. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia (2006) 313–320
42. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of the 17th International Conference on Computational Linguistics, Montreal, Canada (1998) 768–774
43. Hamp, B., Feldweg, H.: GermaNet – a Lexical-Semantic Net for German. In: Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications, Madrid, Spain (1997) 9–15
44. Kunze, C.: Extension and Use of GermaNet, a Lexical-Semantic Database. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece (2000) 999–1002
45. Lemnitzer, L., Kunze, C.: Computerlexikographie. Gunter Narr Verlag, Tübingen, Germany (2007)
46. Fellbaum, C., ed.: WordNet – An Electronic Lexical Database. Language, Speech, and Communication. MIT Press, Cambridge, MA (1998)
47. Faaß, G., Heid, U., Schmid, H.: Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta (2010) 803–810
48. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation **43**(3) (2009) 209–226
49. Schiller, A., Teufel, S., Stckert, C., Thielen, C.: Guidlines für das Tagging deutscher Textcorpora mit STTS. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, and Seminar für Sprachwissenschaft, Universität Tübingen. (1999)
50. Miller, G.A., Fellbaum, C.: Semantic networks of english. Cognition **41** (1991) 197–229
51. Church, K.W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics **16**(1) (1990) 22–29
52. Christopher D. Manning, P.R., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)